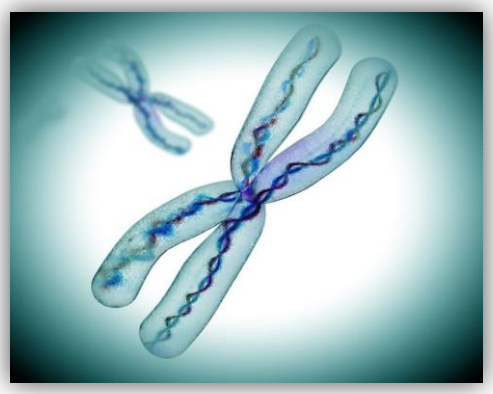


Influence of repetitive elements on pathogenic copy number variants (CNVs) associated with X-Linked Intellectual Disability (XLID)

Ana Rita Pereira Cardoso

Mestrado em Genética Forense

Departamento de Biologia

2016

Orientador

Luísa Azevedo, PhD, Faculdade de Ciências da Universidade do Porto (FCUP), Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Instituto de Investigação e Inovação em Saúde (i3S)

Coorientador

Manuela Oliveira, PhD, Faculdade de Ciências da Universidade do Porto (FCUP), Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Instituto de Investigação e Inovação em Saúde (i3S)

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

"Courage is resistance to fear, mastery of fear - not absence of fear."

Mark Twain

Acknowledgements

Throughout this academic journey I was inspired by several people to whom I owe the successful closure of this chapter and therefore deserve to be rightfully acknowledged.

I would like first to thank my supervisor Luísa Azevedo for encouraging me to expand my creativity, explore a completely novel topic and work independently. This thesis was built from scratch and ever since we started you have always respected my ambition to work on a topic related to clinical genetics. Your supervision helped me to grow as a scientist and to attain new professional perspectives.

I would also like to thank my co-supervisor Manuela Oliveira for contributing to my work by giving new ideas and improvement suggestions. I appreciate that you have reminded me of very important details that dictated the quality of the work.

To my dear parents, António and Manuela, thank you very much for being great parents by supporting and motivating my choices and by raising me to be independent, persistent and ambitious. Words cannot describe the amount of admiration and pride I feel towards both of you. I mostly dedicate this work to you because your emotional and financial supports were utmost essential at pursuing my master's degree.

To my dear boyfriend and classmate Joel, I wish to express my deepest gratitude for all the support, motivation, patience and goofy moments you provided me with throughout this year. Your presence has surely helped me to accomplish this goal and I wish you all the luck with your master thesis and professional career.

I wish to thank my closest classmates Diana, Cátia, Letícia, Jenni, Andreia, and Zita for all the joyful moments and exchanging of ideas during our lunches at FCUP, planetarium, and "Já Lá Foste". I wish you all the best and hope that very bright futures await you.

Last but not least, I appreciate the support of my closest family, Patrícia, André, Clara, António, Olinda and Victor. You are a constant presence in my life and the regular exchanging of ideas during conversations has surely influenced my train of thought.

Abstract

Copy number variants (CNVs) are genomic structural lesions with more than one kilobase. They are considered polymorphisms and used as molecular markers when observed at a 1% frequency in populations. CNVs are alterations that involve duplication, deletion or triplication of certain genes or genomic portions, and this is the reason why CNVs have been associated with several genetic disorders and syndromes. Repetitive elements in the genome (e.g.: LCRs, LINEs, SINEs, etc.) generate instability of the genome and influence the formation of these variants; therefore, it is essential to study their involvement in pathogenic CNV generation.

This work involved the *in silico* analysis of the flanking regions of the breakpoints of pathogenic CNVs associated with the X-linked intellectual disability (XLID) phenotype. The analysis involved a research for XLID genes that were primarily associated with defective neurodevelopment and X-linked mental retardation (MRX). The selected genes were then researched in a mutational database to retrieve reported pathogenic copy number variants that matched the desired phenotype. The breakpoints of the pathogenic CNVs were then repeat masked to estimate the proportion of repetitive elements in their vicinities. Additionally, features such as sequence size and distance between pathogenic and non-pathogenic reported CNV breakpoints were investigated.

Overall, it is clear that both low copy repeats and retrotransposons are major contributors to genomic instability. Fifty-one pathogenic copy number variants in the X chromosome were found to be associated with a XLID phenotype. Retrotransposons such as LINEs and SINEs were found to be frequent in the flanking regions of pathogenic CNVs. Some pathogenic breakpoints overlap a repetitive sequence and thus could be under the influence of local genomic architecture. Furthermore, a significant discrepancy between sequence size was noticed between pathogenic CNVs and non-pathogenic CNVs. Both types of CNVs tend to be proximal; however, pathogenic CNVs are considerably larger, affecting more genes and regulatory sequences than non-pathogenic CNVs.

Keywords: CNV, pathogenic, repetitive elements, structural variant, intellectual disability, X chromosome

Resumo

Copy number variants (CNVs) são mutações estruturais presentes no genoma com mais de 1000 nucleotídeos. São considerados polimorfismos e usados como marcadores moleculares quando observados com uma frequência de 1% nas populações. São alterações que englobam a duplicação, deleção ou triplicação de determinados genes ou porções genómicas e, por este motivo, podem causar graves doenças e síndromes genéticos. Elementos repetitivos no genoma (ex: LCRs, LINEs, SINEs, etc.) geram instabilidade e influenciam a formação destas variantes, pelo que o estudo do seu envolvimento na origem dos CNVs patogénicos é essencial.

Este trabalho envolveu uma análise *in silico* às regiões flanqueantes dos limites de CNVs patogénicos associados com fenótipo de atraso mental ligado ao cromosoma X (XLID). A análise envolveu a pesquisa de genes XLID primariamente associados a defeitos de neurodesenvolvimento e atraso mental associado ao cromossoma X (MRX). Os genes selecionados foram posteriormente sujeitos a pesquisa numa base de dados mutacional de modo a detetar CNVs patogénicos associados ao fenótipo pretendido. Os *breakpoints* dos CNVs patogénicos foram sujeitos a deteção de elementos repetitivos de modo a estimar a proporção de elementos nas regiões flanqueantes dos CNVs. Adicionalmente, parâmetros como tamanho de sequência e distância entre *breakpoints* de CNVs patogénicos e não patogénicos foram avaliados.

No geral, os *low copy repeats* e os retrotransposões contribuem em grande parte para a instabilidade genómica. Mais de cinquenta variações estruturais patogénicas no cromosoma X foram associadas ao fenótipo XLID. Retrotransposões como LINEs e SINEs foram os elementos que se revelaram mais abundantes nas regiões flanqueantes dos CNVs patogénicos. Alguns *breakpoints* patogénicos pertencem à parte da sequência de um elemento repetitivo. Adicionalmente, observou-se uma discrepância significativa entre o tamanho dos CNVs patogénicos e não patogénicos. Apesar de haver proximidade entre os dois tipos de variantes, os CNVs patogénicos englobam uma maior parte da sequência genómica, consequentemente afetando mais genes e sequências de regulação genómicas.

Palavras-chave: CNV, patogénico, elementos repetitivos, variante estrutural, atraso mental, cromossoma X

List of Tables

List of Figures

List of Abbreviations

- **BIR:** Break-Induced Replication
- **Bp:** Base-pair
- **CNV:** Copy Number Variant
- **DNA:** Deoxyribonucleic Acid
- **ERV:** Endogenous Retrovirus
- **ERVL:** Endogenous Retrovirus-Like
- **ERVL-MaLR:** Endogenous Retrovirus-Like / Mammalian-apparent Long Terminal Repeat
- **FoSTeS:** Fork Stalling and Template Switching
- **hAT-Charlie:** hobo/Ac/Tam3-Charlie
- **HERV:** Human Endogenous Retrovirus
- **Kb:** Kilo-base-pair
- **LCR:** Low Copy Repeat
- **LINE:** Long Interspersed Nuclear Element
- **LTR:** Long Terminal Repeat
- **MaLR:** Mammalian-apparent Long Terminal Repeat
- **Mbp:** Mega-base-pair
- **MIR:** Mammalian-wide Interspersed Repeat
- **MRX:** Mental Retardation, X-linked
- **mtDNA:** Mitochondrial DNA
- **NAHR:** Non-Allelic Homologous Recombination
- **NHEJ:** Non-Homologous End Joining
- **RNA:** Ribonucleic Acid
- **SD:** Segmental Duplication
- **SINE:** Short Interspersed Nuclear Element
- **SNP:** Single Nucleotide Polymorphism
- **STR:** Short Tandem Repeat
- **TcMar-Tigger:** Tc1/Mariner-Tigger
- **TE:** Transposable Element
- **TIR:** Terminal Inverted Repeat
- **XLID:** X-Linked Intellectual Disability

Table of Contents

Introduction

1. Copy number variants

Copy number variants (CNVs) are genomic sequences of large size (from more than 1 kb to several Mb) that constitute one of the major sources of variation in the human genome [1-3]. CNVs are structural variants and differ regarding the type of variation (gains or losses), genomic position and length [4]. Additionally, these variants represent changes in gene copy number that may disrupt gene structure and regulation; therefore, a gene associated with a specific phenotype may be present in a single copy in the genome of an individual while present in several copies in a different genome [5, 6].

The variation of the copy number of a DNA segment can be non-pathogenic or pathogenic [6]. For instance, non-pathogenic variation among individuals includes distinct patterns of gene expression associated with physical features (e.g., height and body mass index), nutrition, olfactory receptors, immunologic responses, drug metabolism, and hormonal dosage [2, 4, 7]. As such, non-pathogenic variation can be associated with distinct degrees of adaptation among individuals [4, 6, 8]. On the other hand, pathogenic variation includes aberrant lesions involved in diseases such as rheumatoid arthritis, Crohn's disease, asthma, nephronophthisis, azoospermia, psoriasis, Parkinson's disease, neurodevelopmental syndromes (e.g., autism, intellectual disability, schizophrenia, etc.) and cancer [7, 9-14].

1.1. Population genetics

Copy number variants are present in the normal population and due to their size they considerably affect more nucleotides per genome than single nucleotide polymorphisms (SNPs), contributing to diverse phenotypic traits [6, 8]. Depending on the CNV frequency and local SNP density, CNVs can be tagged by nearby SNPs and used as a parameter for studies involving linkage disequilibrium among human populations [15].

1.2. Forensic genetics

Recent studies involving global copy number variants underlined their importance and utility as genetic markers. Even though CNVs have a high mutation rate, through evolutionarily ancient CNVs fixed in populations it is possible to estimate global migration routes in both female and male lineages [6]. Therefore, CNV analysis could be used as a complementary method in Y chromosome haplogroup analysis and mitochondrial DNA (mtDNA) SNP analysis.

Additionally, CNVs are important in forensic genetics since they might interfere with genetic profiling involving short tandem repeats (STRs). For instance, CNV gains may lead to an aberrant tri-allelic pattern at the TPOX locus, which is a widely used STR marker [16]. On the other hand, a deleterious CNV may give rise to a null allele for a given STR locus and consequently be the leading cause for a monosomic STR profile [17].

1.3. Molecular mechanisms and repetitive elements

The breakpoints of copy number variants are classified as recurrent or non-recurrent. Recurrent rearrangements share the size of the variant, have clustered breakpoints and are carried by unrelated individuals, whereas non-recurrent variants vary in size, have scattered breakpoints and occur sporadically [10, 18]. It is assumed that there are significant differences between the molecular mechanisms underlying the formation of these two types of breakpoints.

The mechanisms originating CNV are extensive and differ regarding complexity and basic mechanistic properties. Simple recurrent genomic copy gains or losses may result from a single recombination event between highly homologous repeats whereas non-recurrent CNVs may involve several segments and result from more complex mechanisms. These mechanisms may not depend on high sequence homology, such as non-homologous end joining (NHEJ), fork stalling and template switching (FoSTeS) and break-induced replication (BIR) (Fig. 1) [18].

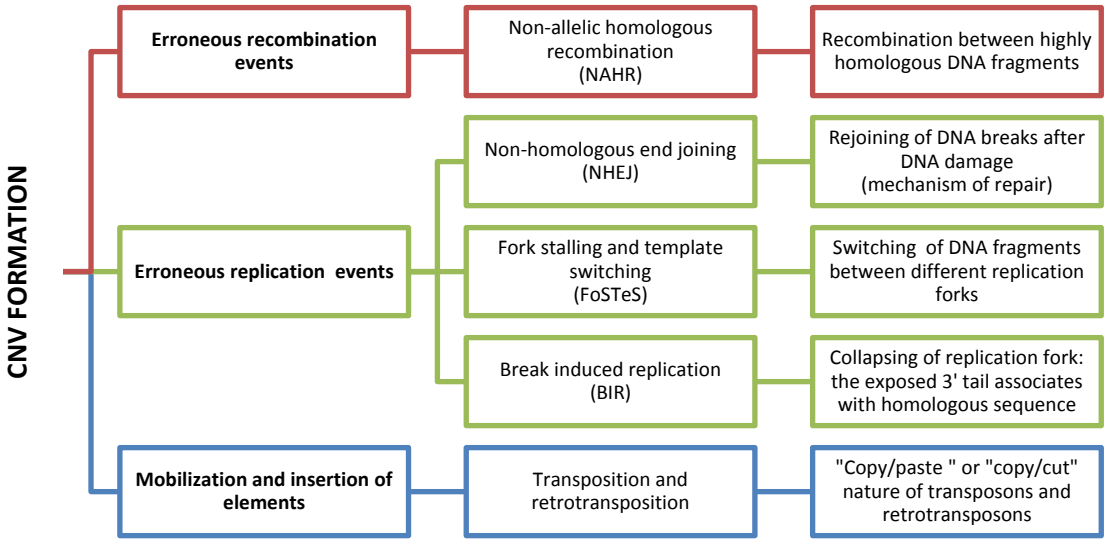


Fig.1 – Mechanisms of copy number variant formation. Adapted from references [18-20].

Recombination-induced errors during mitosis or meiosis are the most frequent cause of CNV formation. The process of recombination between misaligned directly oriented homologous sequences induces the formation of deletions and duplications that may negatively affect gene dosage and trigger genomic disorders (Fig. 2) [19, 21].

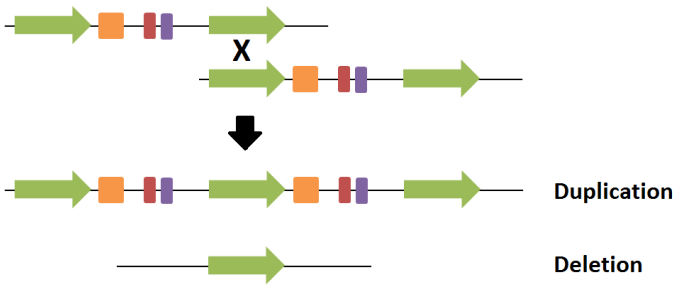


Fig. 2 –Schematic representation of duplication and deletion of genes (orange, red and purple blocks) through non-allelic homologous recombination (NAHR) between directly oriented homologous repeats (green arrows) in misaligned chromatids.

1.3.1. Low copy repeats

The simplest and major mechanism underlying CNV formation is non-allelic homologous recombination (NAHR) between highly homologous low copy repeats (LCRs) [5, 10, 18, 19, 21-24]. These repeats are defined as paralogous segments with 10-400 kb in size and 95-97% of identity which confer genomic instability and are associated with structural rearrangements that cause genomic disorders [10, 18, 19, 25]. However, there are some requirements for the occurrence of NAHR events between LCRs, as displayed in Fig. 3.

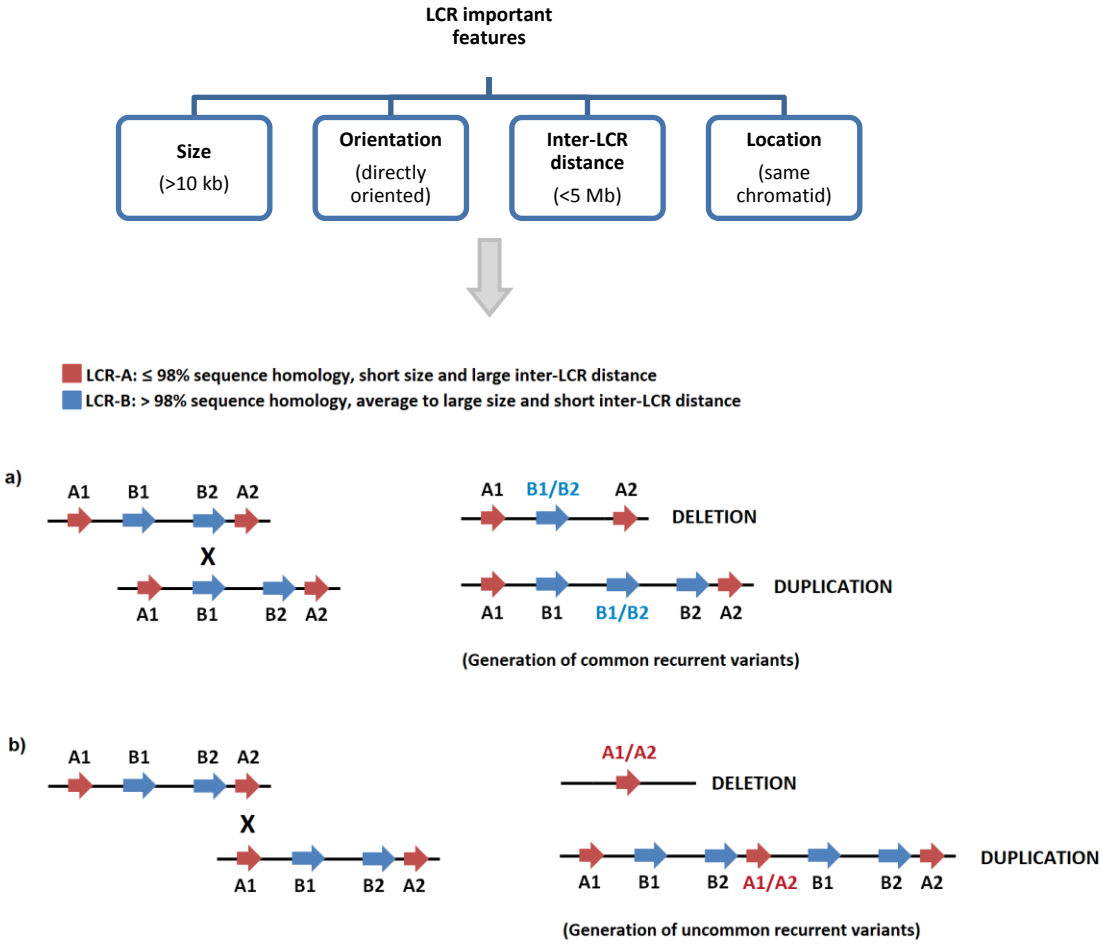


Fig. 3 – Optimal LCRs features for the occurrence of NAHR events that lead to CNV formation. Distinct LCR pairs with counter features such as homology, size and inter-LCR distance influence NAHR rate and lead to the formation of common recurrent **(a)** or uncommon recurrent **(b)** copy number variants. A1/A2 and B1/B2 represent recombinant LCRs. Both types of LCR are directly oriented. Adapted from Cardoso et al. (2016).

Previous studies underlined the importance of LCR length (positively correlated with NAHR rate) and the inter-LCR distance, which inversely affects NAHR events [10,

18, 19, 21]. As such, common recurrent copy number variants originate from NAHR events between directly oriented long LCRs with a high degree of homology and short inter-LCR distances (Fig. 3a). On the contrary, rare recurrent CNVs are generated through NAHR events between LCRs with exact opposite features (Fig. 3b).

A particular subclass of abundant and shorter LCRs includes segmental duplications which are segments with 1-100 kb in size that share 90-95% of nucleotide similarity [10, 19]. These elements are also frequently involved in NAHR events and CNV formation.

1.3.2. Transposable elements

NAHR events that lead to CNV formation may also involve high copy repeats, a different class of repetitive sequences that include interspersed repeats which constitute about 44-45% of the human genome [19, 26].

Mobile or transposable elements (TEs) are important interspersed repeats that move to different regions in the genome through a copy-paste mechanism [20]. A particularly abundant class of TEs includes elements that encode a reverse transcriptase and transpose via a RNA intermediate (retrotransposons) whereas a second class relies on excision and reintegration via a DNA intermediate (DNA transposons) [27-29].

Moreover, the presence or absence of a specific 100 nucleotide-repeated sequence (long terminal repeat, LTR) flanking the TE distinguishes two subclasses of retrotransposons (Fig. 4) [27, 29-31].

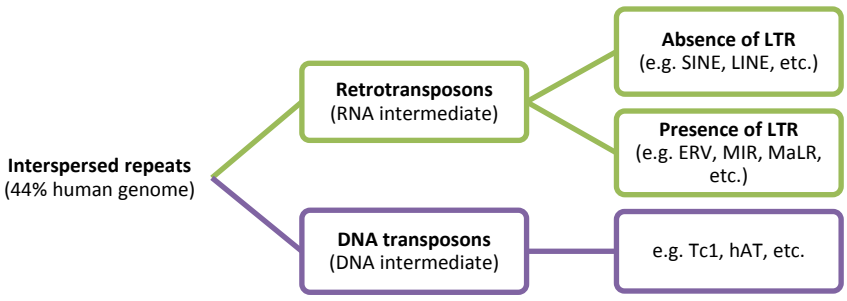


Fig. 4 – Classification of interspersed repeats[1].

[1] LTR – Long terminal repeat; SINE – Short interspersed nuclear element; LINE – Long interspersed nuclear element; ERV – Endogenous retrovirus; MIR – Mammalian-wide interspersed repeat; MaLR – Mammalian-apparent LTR; hAT – hobo/Ac/Tam3.

The structure of most LTR retrotransposons allows autonomous mobilization by controlling transcription and integration in the genome, whereas non-LTR retrotransposons and DNA transposons may or may not be autonomous TEs [20, 30]. For instance, long interspersed repeats (LINEs) are autonomous TEs while short interspersed repeats (SINEs) depend on LINEs' structure to successfully transpose (Table 1).

Table 1 – Characteristics of the most abundant and best described classes of retrotransposons and DNA transposons in humans. Adapted from references [5, 19, 20, 27, 28, 30-32].

TE class	Name	Human genomic percentage	Size (kb)	Mechanism	Coding or regulatory sequences	Important elements
RNA-TEs	Long interspersed nuclear elements (LINEs)	20%	6-8	Autonomous transposition	Endonuclease and reverse transcriptase	L1 (abundant active TE)
	Short interspersed nuclear repeats (SINEs)	11%	0.08-0.4	Dependent on LINEs' machinery (ORF2)	Mobile polymerase III promoter	Alu (abundant active TE)
	Endogenous retroviruses (ERVs)	8%	1.5-10	Autonomous transposition	*gag*, *pol* and *env* (viral genes)	HERV (human specific)
DNA-TEs	Terminal inverted repeats (TIRs)	3%	0.08-3	Copy and paste	Transposase	Tc1/Mariner and hobo/Ac/Tam3 (genetic tools)

In humans, Alu elements and L1 repeats are the most abundant retrotransposons and the only elements that retain retrotransposition activity. However, L1 elements are autonomous whereas Alu repeats are non-autonomous and rely on L1's reverse transcriptase to successfully transpose [33]. Alus are very important regarding clinical genetics since they cause insertional or deleterious mutagenesis, therefore being correlated with genomic disorders such as hemophilia, retinitis pigmentosa and breast cancer [34].

Mammalian LTR retrotransposons are a class of TEs similar to retroviruses regarding structure due to the fact they are fossil sequences of viral infections that occurred throughout mammalian evolution. However, the *env* gene is defective and non-functional [33]. Human endogenous retroviruses (HERVs) are human specific endogenous retroviruses (ERVs) and it has been previously estimated that 12% of the human reference genome could be susceptible to HERV-mediated copy number

variation; a percentage that is higher than the one known for NAHR mediated by LCRs [24]. About 20% of the X chromosome sequence is attributed to susceptibility to HERV-mediated structural rearrangements [24]. Moreover, there are previous reports of pathogenic CNVs occurring due to recombination between HERV elements with 94-95% of homology [18].

The hobo/Ac/Tam3 (hAT) transposons are inactive in humans and constitute about 1.6% of the human genome, therefore being the major category of DNA transposons in humans [35, 36]. The Tc1/Mariner (TcMar) elements are a superfamily of small transposons with about 1-5 kb in length that do not require host mechanisms to successfully transpose [32]. When these elements integrate into the host genome, they insert two nucleotides, whereas hAT elements insert eight nucleotides [28].

2. The X chromosome

In humans, the X chromosome is 155 Mbp long and comprises about 1250 genes, many of which involved in neurodevelopmental pathways and brain function [37, 38]. For this reason, mutations in genes present at various X chromosome loci are likely to induce phenotypic traits associated with neurodevelopmental disorders.

The human X chromosome is considered an interesting case of study regarding copy number variants due to a distinct gene burden between hemizygotic males and heterozygotic females. Therefore, X-linked alleles are more susceptible to positive or negative selection in mammalian males due to their hemizygotic state [39]. Additionally, hemizygosity facilitates the detection of X-linked recessive genetic disorders in male humans [40]. In mammals, the process of X-inactivation in females allows the transcriptional silencing of the genes present in one of the copies to compensate gene dosage and balance gene expression between both sexes [39, 40]. Thus, carrier females may escape the adverse effects of a given variant if it is present in the silenced X chromosome whereas carrier males will always be affected.

2.1. Repetitive elements on X chromosome

The X chromosome is highly enriched with repetitive elements, such as intra-chromosomal segmental duplications and retrotransposons. Interspersed repeats

constitute more than half (56%) of the euchromatic DNA sequence of the X chromosome (155 Mbp) [40].

The most abundant retrotransposons in the X chromosome are LINEs, since L1 elements cover about one-third of the X chromosome [40]. It is suspected that L1 elements are involved in X chromosome inactivation, although this is still a topic of debate between researchers [40, 41].

2.2. X-linked copy number variants

X-linked copy number variants are an interesting research topic due to different modes of inheritance among males and females. A deleterious variant in a male could be lethal due to consequential developmental and functional defects [42].

Assessing the clinical relevance of X-linked CNVs depends on testing the patient's parents. In the case of affected females, the screening of the variant should be performed in both parents whereas in affected males only the mother needs to be tested. When a CNV is maternally inherited, there is the need of an X-inactivation profiling and carrier testing of the mother's family members [43].

2.3. X-linked intellectual disability (XLID)

Intellectual disability is a common and complex neurodevelopmental disorder characterized by an IQ<70 and limitations in intellectual functioning and adaptive behavior [44-47]. This phenotype is associated with genetic predisposition (about 50% of moderate to severe cases) and environmental factors [47]. It has been estimated to affect more males than females (ratio 1.4:1) [46].

Intellectual disability is challenging disorder regarding the definition of the underlying genetic background because multiple genes and different types of mutations may be involved. Additionally, patients may exhibit an isolated ID phenotype (non-syndromic) or ID associated with other symptoms (syndromic), which further emphasizes the degree of complexity of this genetic disorder [47].

It was previously estimated that X-linked CNVs constitute about 10% of the causes of intellectual disability [43]. Since copy number variants affect a higher proportion of the genomic sequence and involve triplication, duplication or deletion of

one or more dosage sensitive genes, it is important to investigate these variants and understand their effect on XLID genes and how they contribute to the phenotypic manifestations observed in affected individuals.

Aims

This work was developed to better understand pathogenic copy number variants and their relationship with repetitive elements. Therefore, the main aims of this work were:

- To perform a literature review on how copy number variants are influenced by local repetitive elements, such as low copy repeats and demonstrate that high copy repeats equally contribute to genomic instability;
- To analyze the breakpoints of pathogenic copy number variants associated with a specific clinical phenotype (X-linked intellectual disability, XLID);
- To perform a comparative analysis between pathogenic (XLID-associated) and non-pathogenic CNVs regarding their genomic proximity and sequence size.

Materials and Methods

Selection of XLID genes on the X chromosome

A comprehensive bibliographic research of X-linked intellectual disability genes was performed to select the genes for which structural or sequence aberrations were mainly associated with the phenotypic trait of intellectual disability. A complete XLID genes list provided by Greenwood Genetic Center [47] in 2015 was used as the main information source on this topic, with additional information from the literature used to further support the work [10, 37, 45, 46, 48-50].

Search for pathogenic copy number variations related to XLID

The 37 selected XLID genes were analyzed in the **DECIPHER v9.10** (***D**atabas**e** of **C**hromosomal **I**mbalance and **P**henotype in **H**umans using **E**nsembl **R**esources*) platform (https://decipher.sanger.ac.uk)[2] [51] to search for associated pathogenic copy number variants. The key parameters considered to the search engine were:

(i) Specific XLID gene;
(ii) Intellectual disability phenotype;
(iii) Three categories of pathogenic CNVs (definitely pathogenic, probably pathogenic and possibly pathogenic).

A total of 47 copy number variants matched the selected criteria. However, some of these variants co-segregated with autosomal copy number variants classified as either benign or pathogenic. Therefore, to avoid overrepresentation of the degree of pathogenicity of XLID-CNVs the pathogenic co-variants (n = 20) were also included in the analysis.

The precise breakpoint coordinates were annotated after redirecting from **DECIPHER v9.10** [51] to the **Ensembl** database (GRCh37/hg19 assembly) [52]. Additionally, a total of 4 pathogenic X-CNVs associated with the intellectual disability

[2] This study makes use of data generated by the **DECIPHER** community. A full list of centres who contributed to the generation of the data is available from http://decipher.sanger.ac.uk and via email from decipher@sanger.ac.uk. Funding for the project was provided by the Wellcome Trust.

phenotype were selected from two different sources [37, 53]. The breakpoints mapped in the NCBI36/hg18 genome version were remapped to the GRCh37/hg19 genome assembly using the **NCBI Genome Remapping Service**. Only remapped coordinates with coverage values ≥ 1.00000 and approximate lengths to the target assembly were considered.

Repeat masking and proportions of repetitive elements

To better characterize the flanking regions of the 51 copy number variations, two segments of 1 kb distal to the breakpoints (both upstream and downstream) was included in the repetitive element analysis, as displayed in Fig. 5. The flanking genomic sequences were retrieved using the **Ensembl** database (GRCh37/hg19 assembly) (http://grch37.ensembl.org/Homo_sapiens) [52] and previously subjected to repeat masking through the **RepeatMasker** service (http://www.repeatmasker.org).

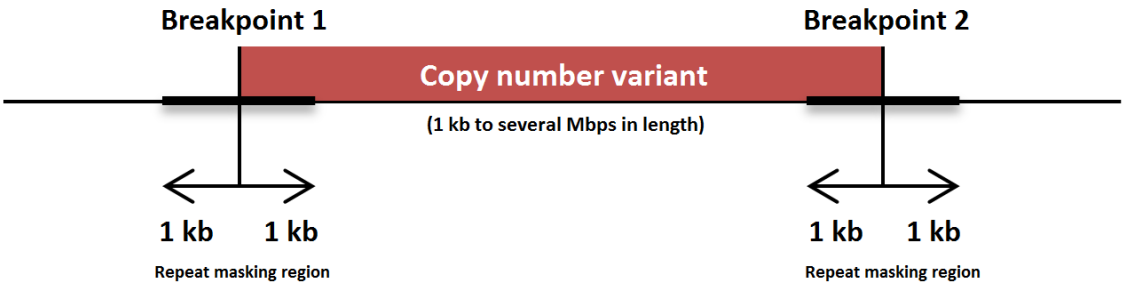


Fig. 5 - Schematic representation of the repeat masked sequence in the flanking regions of pathogenic X-CNVs associated with XLID genes. A 2-kb region flanking each breakpoint was subjected to repeat analysis.

Analysis of repeats overlapping the pathogenic breakpoints

The 51 XLID-associated CNVs and their pathogenic co-variants were subjected to scanning of overlapping repeats in their breakpoints. This analysis was performed in the **UCSC Genome Browser** (GRCh37/hg19 assembly) (https://genome.ucsc.edu) [54].

The data of all the segmental duplications was retrieved from the **UCSC Genome Browser** (GRCh37/hg19 assembly).

Population genetics analysis

A comparative analysis between the breakpoints of the selected 51 XLID-associated CNVs, their 20 pathogenic co-variants and non-pathogenic CNVs was performed using data presented as supplementary materials, provided by Sudmant et al. (2015) [8], in their recent publication on human CNV global diversity and population stratification. Only the most proximal non-pathogenic CNVs were selected for each breakpoint of each pathogenic CNV. Non-pathogenic variants overlapping the core of pathogenic CNVs were only considered if their breakpoints were the most proximal CNV in the vicinity of pathogenic breakpoints.

Results and Discussion

The results in this work are presented in two sections as follows:

Section I: Publications resulting from this thesis

Section II: *In silico* analysis of the breakpoints of XLID-associated pathogenic CNVs

Section I: Publications resulting from this thesis

Although low copy repeats (LCRs) are primarily associated with NAHR events and genomic instability, previous reports suggest that retrotransposons are equally highly influential. Thus, a literature review focused on low copy and high copy repeats influencing the breakpoints of disease-associated CNVs was performed and successfully published in Human Genomics as Cardoso et al. (2016).

Cardoso *et al. Human Genomics* (2016) 10:30
DOI 10.1186/s40246-016-0088-9

Human Genomics

REVIEW Open Access

Major influence of repetitive elements on disease-associated copy number variants (CNVs)

Ana R. Cardoso[1,2,3], Manuela Oliveira[1,2,3], Antonio Amorim[1,2,3] and Luisa Azevedo[1,2,3*]

Abstract

Copy number variants (CNVs) are important contributors to the human pathogenic genetic diversity as demonstrated by a number of cases reported in the literature. The high homology between repetitive elements may guide genomic stability which will give rise to CNVs either by non-allelic homologous recombination (NAHR) or non-homologous end joining (NHEJ). Here, we present a short guide based on previously documented cases of disease-associated CNVs in order to provide a general view on the impact of repeated elements on the stability of the genomic sequence and consequently in the origin of the human pathogenic variome.

Keywords: Copy number variants (CNVs), Genetic diseases, Genomic structural variation, Low copy repeats, Retrotransposons, LINE, SINE, Non-allelic homologous recombination (NAHR)

Background

Copy number variants (CNVs) are structural genomic markers (insertions or deletions) ranging in size from 1 kb to several megabytes for each copy. They are categorized as copy number polymorphisms (CNPs) when multiple allelic states exist in the population or as rare copy number variants when they are found to be associated with genetic diseases (pathogenic copy number variants) [1, 2]. The origin of each repeated element of the CNV is influenced by the local genomic architecture which includes the presence of repetitive sequences within or flanking the repeated segment [3–7]. These repeated sequences drive non-allelic homologous recombination (NAHR) events which result in recurrent insertions and deletions with similar sequence sizes and clustered breakpoints [3, 6, 8] or non-homologous end joining (NHEJ) events that result in non-recurrent rearrangements that vary in terms of their size and breakpoint location [3, 6, 9]. Although several studies have been demonstrating the contribution of structural variants to the genome architecture, few have specifically focused the influence of repeated sequences at breakpoint locations. With the aim to draw attention to these unstable regions and to establish their role in CNVs, we collated a number of cases of CNV-associated disorders proven to have been generated by low and high copy number repeats which may have influenced the degree of stability of the genomic sequence.

Low copy repeats and their influence on pathogenic CNV formation

Low copy repeats (LCRs) are homologous sequences of ≥1 kb in length which are found in many copies throughout the genome since they are generated by duplication events [3, 10]. Large LCRs (>10 kb) with high sequence homology promote non-allelic homologous recombination (NAHR) [3–6, 10–12] and the misalignment of directly oriented sister chromatids carrying the LCR may promoted NAHR thereby generating both duplications and deletions [4, 5] which in turn give rise to copy number variation. A schematic representation of this process is shown in Fig. 1.

Certain properties of the LCRs such as homology length, sequence similarity, and distance, serve to influence the frequency of NAHR events [3, 6, 12] (Fig. 1). As recently reviewed by Carvalho and Lupski [3], the NAHR

* Correspondence: lazevedo@ipatimup.pt
[1]Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Rua Alfredo Allen 208, 4200-135 Porto, Portugal
[2]IPATIMUP-Institute of Molecular Pathology and Immunology, University of Porto, Rua Júlio Amaral de Carvalho 45, 4200-135 Porto, Portugal
Full list of author information is available at the end of the article

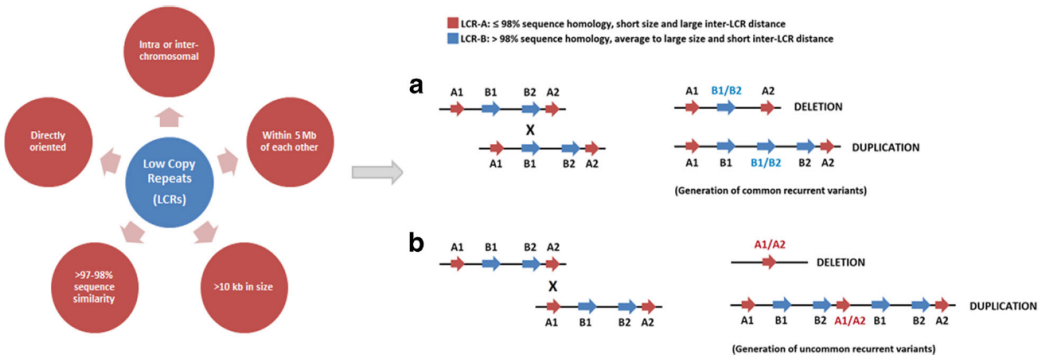


Fig. 1 Optimal LCRs features for the occurrence of NAHR events that result in CNV formation. Distinct LCR pairs with counter features such as homology, size, and inter-LCR distance influence NAHR rate and lead to the formation of common recurrent (**a**) or rare recurrent (**b**) copy number variants. Adapted from [3, 6, 12]

rate varies according to the length of the LCR sequence, the distance between distinct LCR sequences and the DNA sequence. The NAHR rate is, therefore, positively correlated with the LCR length but is inversely proportional to the distance between distinct LCRs [3, 9]. Since there is a high homology between distinct LCR sequences proximal to copy number variation regions; there is also an increased predisposition to NAHR events in these genomic regions [3, 4, 6, 9, 12].

A considerable number of disease-associated CNVs generated by LCRs have been documented and reviewed in previous works (e.g. [3, 6]), but for the purposes of this paper, we have only collated cases for which the specific repetitive element was found at the breakpoints of the structural variant and not those for which the causality of the repeats elements was only suggested. The resulting set is presented in Table 1. For example, a complex array of LCRs spanning a 4-Mb region around the X-linked *MECP2* gene was associated with unique duplications ranging in size from 200 kb to 2.2 Mb in developmentally delayed males [13]. Duplications and deletions affecting the *PLP1* gene causing Pelizaeus-Merzbacher disease (OMIM #312080) are also associated with a specific LCR (LCR-PMD A/B pair) within a 3-Mb region flanking the gene in which a multitude of LCRs are located [14]. LCRs are also frequent at the 2q11-q21.1 locus [11], where recurrent deletions of the *NPHP1* gene (2q13) have been associated with nephronophthisis 1 (OMIM #256100). A 0.3-Mb copy number gain was detected in three X-linked intellectual disability (XLID) families and one sporadic patient [15]. The region overlapped the *GDI1* gene, an important XLID-associated

Table 1 Repetitive elements detected at the breakpoints of CNVs associated with clinical phenotypes

Phenotype	Critical genes	Type of variant	Locus	Repetitive element involved	Ref.
MECP2 duplication syndrome	*MECP2, L1CAM*	Dup	Xq28	Several LCR-MECP2s pairs	[3, 6, 13, 44–46]
Rett syndrome	*MECP2*	Del	Xq28	Several LCR-MECP2 pairs	[6]
Neurofibromatosis type I	*NF1*	Del	17q11.2	NF1-REPs A/B/C	[3, 6]
Nephronophthisis	*NPHP1*	Del	2q13	Several LCR pairs	[11, 47–49]
Mental retardation, X-linked 41 (MRX41)	*GDI1*	Dup/Trip	Xq28	LCR-K1/L2 pair	[15]
Angelman and Prader-Willi syndromes	*UBE3A*	Del	15q11-q13	END-repeats (LCRs)	[6, 50]
Smith-Magenis syndrome	*RAI1* and *PMP22*	Del	17p11.2	SMS-REPs (LCRs)	[3, 6, 18]
Williams-Beuren syndrome	28 dosage-sensitive genes	Dup/Tripe/Del	7q11.23	A/B/C LCR blocks	[3, 6, 51]
15q13.1 microdeletion syndrome	*CHRNA7*	Dup/Trip	15q13.3	BP3/4/5	[3, 6, 52, 53]
3q29 microduplication or microdeletion syndrome	*DLG1, PAK2*	Dup/Del	3q29	A/B/C LCR blocks	[3, 54, 55]
Pelizaeus-Merzbacher disease	*PLP1*	Dup/Del	Xq22	LCR-PMD A/B pair	[3, 6, 14]
DiGeorge syndrome/velo-cardio facial syndrome	*COMT, TBX1*	Del	22q11.2	8 specific LCR22 repeats	[6, 17, 40]
Charcot-Marie-Tooth type 1A	*PMP22*	Dup	17p12	CMTA1-Reps (LCRs)	[3, 6, 37, 38]

gene highly expressed in the brain. The aberration was located in Xq28, a locus that includes other intellectual disability genes and that is frequently associated with recombination events caused by proximal LCRs (e.g., LCR K1/L2). The Angelman syndrome (AS) (OMIM #105830) and Prader-Willi syndrome (PWS) (OMIM #176270) are caused by recurrent 4-Mb deletions at the 15q11-q13 locus. The deleted region is flanked by LCRs [6] and accounts for 70 % of cases of AS and 70–75 % of cases of PWS [16]. The Smith-Magenis syndrome (SMS) (OMIM #182290) results from recurrent deletions of 3.7 Mb at 17p11.2 which account for more than 70 % of cases; about 25 % of affected individuals harbor deletions ranging from 1.5–9 Mb [6, 17]. The deletions are flanked by 200-kb highly homologous LCRs that play a role in generating meiotic NAHR events [16]. These deletions encompass the *RAI1* gene, which is critical in organ and neuronal development—patients with larger deletions manifest a more severe phenotype when the dosage-sensitive gene *PMP22* is deleted [18].

Retrotransposons (high copy repeats) and their influence on pathogenic CNVs

Interspersed repeats are the most common type of high copy repeats, covering about 44 % of the human genome [4]. Retrotransposons account for the majority of transposable elements [5, 7, 19]. These are mobile elements that through reverse transcription have the ability to integrate into different regions [7, 19]. Long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and retrovirus-like elements (LTR transposons) are the three major categories of mammalian retrotransposons (Table 2).

Among LINEs, L1 is the most abundant element, typically of 6–8 kb in length, with the ability to increase genomic instability through NAHR events [4]. It is known that about 83 % of the human genome is prone to LINE-LINE recombination events that contribute to genomic instability and can give rise to unbalanced structural variants [20].

Alu elements are the most common SINEs and have been associated with NAHR events that lead to pathogenic duplications and deletions [3, 4, 21, 22]. Table 3 presents examples of high copy repeats that have been detected at the breakpoints of disease-associated CNVs.

Borun and colleagues [23] reported the presence of CNV breakpoints within Alu elements in the *STK11* gene which lead to the Peutz-Jeghers syndrome (OMIM #175200), where CNVs account for 30 % of cases. The 17p13.3 locus is enriched in copy number variations associated with genomic disorders, such as the Miller-Dieker syndrome (17p13.3 deletion syndrome) (OMIM #247200) and its reciprocal 17p13.3 duplication syndrome (OMIM #613215) [24]. The breakpoints of the reported CNVs at this locus are highly enriched in Alu elements, which mediate these junctions through an Alu-Alu mechanism. About 70 % of CNVs found in the *SPAST* gene have been associated with Alu recombination events [25]. Local Alu-rich architecture predisposes to the formation of pathogenic structural rearrangements associated with spastic paraplegia (OMIM #182601). An extra copy of the *LMNB1* gene at the 5q23 locus has been previously associated with autosomal dominant adult-onset demyelinating leukodystrophy (ADLD) (OMIM #169500). The analysis of twenty ADLD-affected families revealed sixteen duplications ranging from 128 to 475 kb in size, all of them spanning the *LMNB1* gene [26]. The centromeric region of the critical gene is enriched with SINE elements, particularly Alus. Alu-mediated recombination events were also found to be linked to pathogenic deletions at the *OTC* gene [27], a urea cycle gene for which a significant number of structural variants are known [28]. NAHR events between Alu repeats are also strongly correlated with the birth of structural rearrangements at the Alu-rich *BRCA1* locus [29] which is associated with breast cancer. Duplications (220 to 394 kb) and a triplication (1.61 to 2.04 Mb) of the *SNCA* gene located at 4q21 locus have been implicated in autosomal dominant Parkinson's disease (PD1 and PD4) (OMIM #168601, #605543). The phenotypic severity is consistent with a gene dosage effect [6]. Regarding recessive PD (OMIM #600116), about one third of pathogenic variants associated with the *PRKN* gene are CNVs occurring between exon 2 and exon 5, which may therefore be considered to be a

Table 2 Main characteristics of the most abundant retrotransposons [4, 5, 7, 19]

	Retrotransposons (interspersed repeats)—44 % human genome		
	Non-long terminal repeat (LTR)		Long terminal repeat (LTR)
Repetitive element	Long interspersed nuclear repeats (LINEs)	Short interspersed nuclear repeats (SINEs)	Endogenous retroviruses (ERV)
Genomic coverage	20 %	11 %	8 %
Features	• L1 is the most abundant class • Autonomous transposons • Reverse transcriptase (RT) encoded by LINE-1	• Alu is the most abundant class • Dependent on LINEs transposable machinery • Mobile polymerase III promoter • 100–400 bp in length	• Reduced transposable activity • Presence of *gag* and *pol* viral genes

Table 3 High copy repeats detected at the breakpoints of CNVs associated with clinical phenotypes

Phenotype	Critical genes	Type of variant	Locus	Repetitive elements involved	Ref.
Peutz-Jeghers syndrome	*STK11*	Del	19p13.3	Several AluY/AluY pairs	[23]
Spastic paraplegia 4	*SPAST, SLC30A6*	Dup/Del	2p22.3	Several Alu pairs	[25]
OTC deficiency	*OTC*	Del	Xp11.4	AluSx/AluSq pair	[27, 56]
Miller-Dieker syndrome and 17p13.3 duplication syndrome	*LIS1*	Del	17p13.3	Several Alu pairs	[6, 24]
Breast cancer	*BRCA1*	Del	17q21.31	AluSx/AluSc pair	[29, 57]
Autosomal dominant adult-onset demyelinating leukodystrophy (ADLD)	*LMNB1*	Dup/Trip	5q23.2	LIPA3 LINE repeats AluYA/AluYB pair	[26]
Azoospermia	*AZFa*	Del	Yq11	HERV15 A/B proviruses	[34, 58]
Mental retardation, X-linked 60 (MRX60)	*OPHN1*	Del	Xq12	AluY/AluY pair	[35]
Pelizaeus-Merzbacher disease	*PLP1*	Del	Xq22	AluSq/AluSx pair	[3, 6, 59]
DiGeorge syndrome/velo-cardio facial syndrome	*COMT, TBX1*	Del	22q11.2	Unclassified Alu/Alu pair	[6, 17, 40]
Charcot-Marie-Tooth type 1A	*PMP22*	Dup	17p12	AluY/AluY pair AluSg/AluSg pair	[39]
Williams-Beuren syndrome	28 dosage-sensitive genes	Dup/Del	7q11.23	AluS subfamily elements	[36]
Parkinson's disease	*SNCA*	Dup/Trip	4q21	Several Alu pairs	[32]

recombination hotspot [30, 31]. Ross and colleagues [32] reported the presence of Alu and LINE1 elements at the *SNCA* locus that may contribute to the genomic instability at this locus.

Human endogenous retroviruses (HERVs) represent about 4.9 % of the human genome [4]. Sequences with about 95 % sequence similarity were previously associated with NAHR events and recurrent CNVs, some of which with pathogenic implications [3, 33]. For example, the occurrence of NAHR between a particular set of HERV elements flanking the male fertility *AZFa* locus in the Y chromosome is strongly associated with pathogenic deletions associated with male infertility (OMIM #400042, #415000) [4, 34].

Pathogenic copy number variants associated with both LCRs and retrotransposons

The breakpoints of some disease-associated CNVs have been reported to be caused by more than one type of repetitive elements which indicates that the same phenotype involves both low copy and high copy repeats that affect the stability of a target gene. Bergmann and colleagues [35] conducted a family study in which five brothers shared the same phenotypic pattern that included intellectual disability. The analysis of the OPHN1 locus (Xq12) revealed the presence of a 17.6-kb intronic deletion and the breakpoints spanning the deletion revealed two highly homologous Alu repeats and additional repetitive sequences (interspersed and simple repeats).

A recurrent deletion of 1.6 to 1.8 Mb (>95 % of the patients) at the 7q11.23 locus causes the Williams-Beuren syndrome (OMIM #194050) [6]. Genes within this region are dosage-sensitive and the recurrently deleted region encompasses a total of 28 genes. This locus is characterized by highly homologous flanking LCRs that contribute to NAHR events [6]. Antonell and colleagues [36] reported the presence of Alu elements at the junctions of large duplicated blocks in 7q11.23 suggesting the influence of these retrotransposons in the generation of large LCRs.

Heterozygous duplication and reciprocal deletions of a 1.4–1.5-Mb segment at the 17p12 locus have been previously linked with the Charcot-Marie-Tooth type 1A syndrome (CMT1A) (OMIM #118220). About 70 % of CMT1A patients have a recurrent duplication of the dosage-sensitive PMP22 locus and the NAHR event that gave rise to this copy number variation was mediated by LCRs [3, 6, 37, 38]. A study by Zhang and colleagues [39] revealed the presence of SINEs (Alu elements) and LINEs (L1 and L2) as well as LCRs within the breakpoints of rare non-recurrent deletions and duplications at the CMT1A locus.

About 96 % of the DiGeorge syndrome (DGS) (OMIM #188400)- and velo-cardio-facial syndrome (VCFS) (OMIM #192430)-affected patients harbor a 1.5–3 Mb deletion at the 22q11.2 locus that includes 24 to 30 genes [16]. The breakpoints of the common recurrent deletions at this locus are associated with LCRs [17] and one Alu sequence [40]. Both the deletions and duplications at this locus are generated by NAHR events

between the repeated regions flanking the CNV, specifically the low copy repeat known as LCR22 [41]. Furthermore, 20–25 % of individuals who harbor this deletion also show signs of schizophrenia, mood disorders, and other behavioral alterations [41].

Conclusions

Although the majority of genetic diseases are caused by non-structural variants (e.g. [42, 43], an increasing number of causative mutations have been associated with CNVs and these cases were the focus of this short review. Low copy repeats and retrotransposons are the major contributors to CNV formation. Recurrent CNVs are mainly directed by NAHR events that occur between highly homologous LCR sequences. In terms of non-recurrent CNVs, NHEJ (among other molecular mechanisms [3]) generally occurs between sequences with a degree of homology lower than that observed between distinct LCRs. The diversity of breakpoint junctions of non-recurrent variants renders the establishment of phenotype-genotype relationships less reliable because the sequence that is deleted or duplicated in each patient is different and the affected region may also involve other genes. This review focused on disease-associated CNVs in order to show that although numerous cases of instability driven by repeated sequences around the affected locus (or loci) have been documented, we are still far from understanding all the phenotypic complexities associated with these unbalanced variants, mainly because the number of reported cases is still too small to draw general conclusions. Finally, it is important to mention that collated data, such as those presented in this paper, pertaining to the pathogenic structural variome are expected to drive future studies with the aim of establishing a map of unstable genomic hotspots which promises to be useful in the context of clinical genetic testing where the determination of the molecular basis of Mendelian and complex diseases (e.g., cancer) is of paramount importance.

Abbreviations

CNV: Copy number variant; LCR: Low copy repeat; CNP: Copy number polymorphism; LINE: Long interspersed nuclear repeat; SINE: Short interspersed nuclear repeat; LTR: Long terminal repeat; NAHR: Non-allelic homologous recombination; NHEJ: Non-homologous end joining; XLID: X-linked intellectual disability

Acknowledgements

Not applicable.

Funding

IPATIMUP integrates the i3S Research Unit, which is partially supported by FCT, the Portuguese Foundation for Science and Technology. This work is funded by FEDER funds through the Operational Programme for Competitiveness Factors - COMPETE and National Funds through the FCT-Foundation for Science and Technology, under the projects "PEst-C/SAU/LA0003/2013". M. Oliveira (SFRH/BPD/66071/2009) was supported by FCT fellowships funded by POPH-QREN - Promotion of Scientific Employment, the European Social Fund, and National Funds of the Ministry of Education and Science.

Availability of data and materials

Not applicable.

Authors' contributions

ARC performed literature search. ARC and LA drafted the manuscript. MO and AA discussed the data and performed critical revisions to the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Rua Alfredo Allen 208, 4200-135 Porto, Portugal. [2]IPATIMUP-Institute of Molecular Pathology and Immunology, University of Porto, Rua Júlio Amaral de Carvalho 45, 4200-135 Porto, Portugal. [3]Department of Biology, Faculty of Sciences, University of Porto, Rua do Campo Alegre S/N, 4169-007 Porto, Portugal.

Received: 19 July 2016 Accepted: 16 September 2016
Published online: 23 September 2016

References

1. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, et al. Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet. 2009;84(2):148–61.
2. Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. Hum Mol Genet. 2009;18(R1):R1–8.
3. Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17(4):224–38.
4. Chen L, Zhou W, Zhang L, Zhang F. Genome architecture and its roles in human copy number variation. Genomics Inform. 2014;12(4):136–44.
5. Jobling MA, Hollox E, Hurles M, Kivisild T, Tyler-Smith C. Human evolutionary genetics: origins, peoples and disease. 2nd ed. New York: Garland Science; 2014.
6. Lee JA, Lupski JR. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. Neuron. 2006;52(1):103–21.
7. Strachan T, Read A. Human Molecular Genetics. 4th ed. New York:Garland Science; 2011.
8. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. Pathogenetics. 2008;1(1):4.
9. Liu P, Lacaria M, Zhang F, Withers M, Hastings PJ, Lupski JR. Frequency of nonallelic homologous recombination is correlated with length of homology: evidence that ectopic synapsis precedes ectopic crossing-over. Am J Hum Genet. 2011;89(4):580–8.
10. Potier MC, Golfier G, Eichler EE. Chromosome-specific repeats (Low-copy Repeats). In:eLS. Chichester: Wiley; 2007.
11. Dittwald P, Gambin T, Szafranski P, Li J, Amato S, Divon MY, et al. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. Genome Res. 2013;23(9):1395–409.
12. Peng Z, Zhou W, Fu W, Du R, Jin L, Zhang F. Correlation between frequency of non-allelic homologous recombination and homology properties: evidence from homology-mediated CNV mutations in the human genome. Hum Mol Genet. 2015;24(5):1225–33.
13. del Gaudio D, Fang P, Scaglia F, Ward PA, Craigen WJ, Glaze DG, et al. Increased MECP2 gene copy number as the result of genomic duplication in neurodevelopmentally delayed males. Genet Med. 2006;8(12):784–92.
14. Lee JA, Inoue K, Cheung SW, Shaw CA, Stankiewicz P, Lupski JR. Role of genomic architecture in PLP1 duplication causing Pelizaeus-Merzbacher disease. Hum Mol Genet. 2006;15(14):2250–65.

15. Vandewalle J, Van Esch H, Govaerts K, Verbeeck J, Zweier C, Madrigal I, et al. Dosage-dependent severity of the phenotype in patients with mental retardation due to a recurrent copy-number gain at Xq28 mediated by an unusual recombination. Am J Hum Genet. 2009;85(6):809–22.
16. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet. 2009;84(4):524–33.
17. Shaikh TH, Kurahashi H, Saitta SC, O'Hare AM, Hu P, Roe BA, et al. Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. Hum Mol Genet. 2000;9(4):489–501.
18. Madduri N, Peters SU, Voigt RG, Llorente AM, Lupski JR, Potocki L. Cognitive and adaptive behavior profiles in Smith-Magenis syndrome. J Dev Behav Pediatr. 2006;27(3):188–92.
19. Smit AF. The origin of interspersed repeats in the human genome. Curr Opin Genet Dev. 1996;6(6):743–8.
20. Startek M, Szafranski P, Gambin T, Campbell IM, Hixson P, Shaw CA, et al. Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination. Nucleic Acids Res. 2015;43(4):2188–98.
21. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet. 2002;3(5):370–9.
22. Teixeira-Silva A, Silva RM, Carneiro J, Amorim A, Azevedo L. The role of recombination in the origin and evolution of Alu subfamilies. PloS one. 2013;8(6):e64884.
23. Borun P, De Rosa M, Nedoszytko B, Walkowiak J, Plawski A. Specific Alu elements involved in a significant percentage of copy number variations of the STK11 gene in patients with Peutz-Jeghers syndrome. Fam Cancer. 2015;14(3):455–61.
24. Gu S, Yuan B, Campbell IM, Beck CR, Carvalho CM, Nagamani SC, et al. Alu-mediated diverse and complex pathogenic copy-number variants within human chromosome 17 at p13.3. Hum Mol Genet. 2015;24(14):4061–77.
25. Boone PM, Yuan B, Campbell IM, Scull JC, Withers MA, Baggett BC, et al. The Alu-rich genomic architecture of SPAST predisposes to diverse and functionally distinct disease-associated CNV alleles. Am J Hum Genet. 2014;95(2):143–61.
26. Giorgio E, Rolyan H, Kropp L, Chakka AB, Yatsenko S, Di Gregorio E, et al. Analysis of LMNB1 duplications in autosomal dominant leukodystrophy provides insights into duplication mechanisms and allele-specific expression. Hum Mutat. 2013;34(8):1160–71.
27. Quental R, Azevedo L, Rubio V, Diogo L, Amorim A. Molecular mechanisms underlying large genomic deletions in ornithine transcarbamylase (OTC) gene. Clin Genet. 2009;75(5):457–64.
28. Azevedo L, Stolnaja L, Tietzeova E, Hrebicek M, Hruba E, Vilarinho L, et al. New polymorphic sites within ornithine transcarbamylase gene: population genetics studies and implications for diagnosis. Mol Genet Metab. 2003; 78(2):152–7.
29. Ewald IP, Ribeiro PL, Palmero EI, Cossio SL, Giugliani R, Ashton-Prolla P. Genomic rearrangements in BRCA1 and BRCA2: a literature review. Genet Mol Biol. 2009;32(3):437–46.
30. Toft M, Ross OA. Copy number variation in Parkinson's disease. Genome Med. 2010;2(9):62.
31. Hedrich K, Eskelson C, Wilmot B, Marder K, Harris J, Garrels J, et al. Distribution, type, and origin of Parkin mutations: review and case studies. Mov Disord. 2004;19(10):1146–57.
32. Ross OA, Braithwaite AT, Skipper LM, Kachergus J, Hulihan MM, Middleton FA, et al. Genomic investigation of alpha-synuclein multiplication and parkinsonism. Ann Neurol. 2008;63(6):743–50.
33. Campbell IM, Gambin T, Dittwald P, Beck CR, Shuvarikov A, Hixson P, et al. Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination. BMC Biol. 2014;12:74.
34. Bosch E, Jobling MA. Duplications of the AZFa region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility. Hum Mol Genet. 2003;12(3):341–7.
35. Bergmann C, Zerres K, Senderek J, Rudnik-Schoneborn S, Eggermann T, Hausler M, et al. Oligophrenin 1 (OPHN1) gene mutation causes syndromic X-linked mental retardation with epilepsy, rostral ventricular enlargement and cerebellar hypoplasia. Brain. 2003;126(Pt 7):1537–44.
36. Antonell A, de Luis O, Domingo-Roura X, Perez-Jurado LA. Evolutionary mechanisms shaping the genomic structure of the Williams-Beuren syndrome chromosomal region at human 7q11.23. Genome Res. 2005;15(9):1179–88.
37. Pentao L, Wise CA, Chinault AC, Patel PI, Lupski JR. Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. Nat Genet. 1992;2(4):292–300.
38. Weterman MA, van Ruissen F, de Wissel M, Bordewijk L, Samijn JP, van der Pol WL, et al. Copy number variation upstream of PMP22 in Charcot-Marie-Tooth disease. Eur J Hum Genet. 2010;18(4):421–8.
39. Zhang F, Seeman P, Liu P, Weterman MA, Gonzaga-Jauregui C, Towne CF, et al. Mechanisms for nonrecurrent genomic rearrangements associated with CMT1A or HNPP: rare CNVs as a cause for missing heritability. Am J Hum Genet. 2010;86(6):892–903.
40. Uddin RK, Zhang Y, Siu VM, Fan YS, O'Reilly RL, Rao J, et al. Breakpoint Associated with a novel 2.3 Mb deletion in the VCFS region of 22q11 and the role of Alu (SINE) in recurring microdeletions. BMC Med Genet. 2006;7:18.
41. Shishido E, Aleksic B, Ozaki N. Copy-number variation in the pathogenesis of autism spectrum disorder. Psychiatry Clin Neurosci. 2014;68(2):85–95.
42. Ferreira F, Esteves S, Almeida LS, Gaspar A, da Costa CD, Janeiro P, et al. Trimethylaminuria (fish odor syndrome): genotype characterization among Portuguese patients. Gene. 2013;527(1):366–70.
43. Quelhas D, Quental R, Vilarinho L, Amorim A, Azevedo L. Congenital disorder of glycosylation type Ia: searching for the origin of common mutations in PMM2. Ann Hum Genet. 2007;71(Pt 3):348–53.
44. Van Esch H. MECP2 duplication syndrome. Mol Syndromol. 2012;2(3-5):128–36.
45. Bauters M, Van Esch H, Friez MJ, Boespflug-Tanguy O, Zenker M, Vianna-Morgante AM, et al. Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. Genome Res. 2008;18(6):847–58.
46. Carvalho CM, Zhang F, Liu P, Patel A, Sahoo T, Bacino CA, et al. Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. Hum Mol Genet. 2009;18(12):2188–203.
47. Saunier S, Calado J, Benessy F, Silbermann F, Heilig R, Weissenbach J, et al. Characterization of the NPHP1 locus: mutational mechanism involved in deletions in familial juvenile nephronophthisis. Am J Hum Genet. 2000; 66(3):778–89.
48. Yuan B, Liu P, Gupta A, Beck CR, Tejomurtula A, Campbell IM, et al. Comparative genomic analyses of the human NPHP1 locus reveal complex genomic architecture and its regional evolution in primates. PLoS Genet. 2015;11(12):e1005686.
49. Konrad M, Saunier S, Heidet L, Silbermann F, Benessy F, Calado J, et al. Large homozygous deletions of the 2q13 region are a major cause of juvenile nephronophthisis. Hum Mol Genet. 1996;5(3):367–71.
50. Amos-Landgraf JM, Ji Y, Gottlieb W, Depinet T, Wandstrat AE, Cassidy SB, et al. Chromosome breakage in the Prader-Willi and Angelman syndromes involves recombination between large, transcribed repeats at proximal and distal breakpoints. Am J Hum Genet. 1999;65(2):370–86.
51. Bayes M, Magano LF, Rivera N, Flores R, Perez Jurado LA. Mutational mechanisms of Williams-Beuren syndrome deletions. Am J Hum Genet. 2003;73(1):131–51.
52. Willatt L, Cox J, Barber J, Cabanas ED, Collins A, Donnai D, et al. 3q29 microdeletion syndrome: clinical and molecular characterization of a new syndrome. Am J Hum Genet. 2005;77(1):154–60.
53. Ballif BC, Theisen A, Coppinger J, Gowans GC, Hersh JH, Madan-Khetarpal S, et al. Expanding the clinical phenotype of the 3q29 microdeletion syndrome and characterization of the reciprocal microduplication. Mol Cytogenet. 2008;1:8.
54. Soler-Alfonso C, Carvalho CM, Ge J, Roney EK, Bader PI, Kolodziejska KE, et al. CHRNA7 triplication associated with cognitive impairment and neuropsychiatric phenotypes in a three-generation pedigree. Eur J Hum Genet. 2014;22(9):1071–6.
55. Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, et al. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet. 2008;40(3):322–8.
56. Shchelochkov OA, Li FY, Geraghty MT, Gallagher RC, Van Hove JL, Lichter-Konecki U, et al. High-frequency detection of deletions and variable rearrangements at the ornithine transcarbamylase (OTC) locus by oligonucleotide array CGH. Mol Genet Metab. 2009;96(3):97–105.
57. Mazoyer S. Genomic rearrangements in the BRCA1 and BRCA2 genes. Hum Mutat. 2005;25(5):415–22.
58. Sun C, Skaletsky H, Rozen S, Gromoll J, Nieschlag E, Oates R, et al. Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. Hum Mol Genet. 2000;9(15):2291–6.
59. Inoue K, Osaka H, Thurston VC, Clarke JT, Yoneyama A, Rosenbarker L, et al. Genomic rearrangements resulting in PLP1 deletion occur by nonhomologous end joining and cause different dysmyelinating phenotypes in males and females. Am J Hum Genet. 2002;71(4):838–53.

Section II: *In silico* analysis of the breakpoints of XLID-associated pathogenic CNVs

Selection of XLID genes on the X chromosome

There are several genes in the current X-linked disability panel. However, only a few are directly involved in the phenotype, systematically associated with mental retardation and tested routinely in laboratories. The present work involved an exhaustive research to select the genes that were more frequently reported to be associated with the intellectual disability phenotype, whether syndromic or non-syndromic. A total of 37 genes (Table 2) are involved in non-syndromic XLID, which indicates that ID is an isolated phenotype and the single clinical manifestation; and syndromic XLID, where the ID phenotype is an unspecific symptom accompanied by other symptoms.

Table 2 – Characteristics of the selected XLID-associated genes [47, 55].

OMIM ref.	Gene	Locus	Designation	Function	XLID-related clinical phenotype[3]
*302910	*CLCN4*	Xp22.2	Chloride channel 4	Chloride transport	MRX49 (NS); MRX15 (NS)
*300828	*PTCHD1*	Xp22.11	Patched domain-containing protein 1	Involvement in hedgehog signaling pathway	MRX; X-linked autism 4
*300075	*RPS6KA3*	Xp22.12	Ribosomal protein S6 kinase	Cell differentiation and maintenance	Coffin-Lowry syndrome; MRX19 (NS)
*300382	*ARX*	Xp21.3	Aristaless-related homeobox	Cerebral development	Partington syndrome; Proud syndrome; MRX29; MRX32; MRX33; MRX38; MRX43; MRX54; MRX76; MRX87 (all NS MRX)
*300206	*IL1RAPL1*	Xp21.3-p21.2	Interleukin-1 receptor accessory protein-like 1	Synaptic regulation	MRX21 (NS)
*300072	*USP9X*	Xp11.4	Ubiquitin-specific protease 9	Regulation of chromosomal alignment	MRX99 (NS)
*300096	*TM4SF2/ TSPAN7*	Xp11.4	Tetraspanin 7	Neurite growth regulation	MRX58 (NS)
*314995	*ZNF41*	Xp11.3	Zinc finger protein 41	Transcription regulation	MRX89 (NS)
*300573	*ZNF674*	Xp11.3	Zinc finger protein 674	Transcription regulation	MRX92 (NS)
*314998	*ZNF81*	Xp11.23	Zinc finger protein 81	Transcription regulation	MRX45 (NS)
*300499	*FTSJ1*	Xp11.23	FTSJ homolog 1	Translation regulation	MRX9 (NS)
*300463	*PQBP1*	Xp11.23	Polyglutamine-binding protein 1	Transcription regulation	Renpenning syndrome 1
*300522	*IQSEC2*	Xp11.22	IQ-motif and SEC7 domain-containing protein 2	Cytoskeletal organization & synaptic function	MRX78 (NS)

[3] S - Syndromic mental retardation / NS – Non-syndromic mental retardation

*300697	*HUWE1*	Xp11.22	HECT, UBA and WWE domain containing 1, E3 ubiquitin protein ligase	Transferase, ligase and binding activity	MRX (Turner type) (S)
*300560	*PHF8/ ZNF422*	Xp11.22	PHD finger protein 8	Histone demethylation	MRX (Siderius type) (S)
*314690	*KDM5C/ JARID1C*	Xp11.22	Lysine-specific demethylase-5C	Transcription repression	MRX (Claes-Jensen type) (S)
*300286	*KLF8/ ZNF741*	Xp11.21	Kruppel-like factor 8	Signalling pathways	MRX (NS)
*300127	*OPHN1*	Xq12	Oligophrenin 1	Cytoskeletal modulation	MRX w/ cerebellar hypoplasia and distinctive facial features (S)
*300189	*DLG3*	Xq13.1	Discs large homolog 3	Synaptic signalling pathways	MRX90 (NS)
*300521	*KIF4A*	Xq13.1	Kinesin family member 4A	Cell division and other cell processes	MRX100 (NS)
*300379	*RLIM*	Xq13.2	Ring finger protein, LIM domain interacting	Co-regulation of transcription factors & chrX inactivation initiation	MRX61 (NS)
*300576	*ZDHHC15*	Xq13.3	Zinc finger DHHC domain-containing protein 15	Receptor activity	MRX91 (NS)
*300553	*BRWD3*	Xq21.1	Bromodomain and WD repeat-containing protein 3	Cell morphology regulation & cytoskeletal organization	MRX93 (NS)
*300460	*PCDH19*	Xq22.1	Protocadherin 19	Cell adhesion	Juberg-Hellman Syndrome
*300204	*MID2*	Xq22.3	Midline 2	Microtubule stabilization	MRX101 (NS)
*300142	*PAK3*	Xq23	p21 protein-activated kinase 3	Signal transduction & cell regulation	MRX30 (NS)
*300157	*ACSL4*	Xq23	Acyl-CoA synthetase long chain family 4	Fatty acids conversion	MRX63 (NS)
*300304	*CUL4B*	Xq24	Cullin 4B	DNA repair mechanisms & protein regulation	MRX (Cabezas type) (S)
*312180	*UBE2A*	Xq24	Ubiquitin-conjugating enzyme E2A	DNA damage repair	MRX (Nascimento type) (S)
*300298	*UPF3B*	Xq24	UPF3 regulator of nonsense transcripts homolog B	Nonsense-mediated decay promotion & translation regulation	MRX14 (NS)
*300395	*THOC2*	Xq25	THO complex, subunit 2	Neuronal development	MRX12 (NS)
*300267	*ARHGEF6*	Xq26.3	Rho guanine nucleotide exchange factor 6	Gene expression, cytoskeletal architecture & apoptosis	MRX46 (NS)
*300104	*GDI1*	Xq28	GDP dissociation inhibitor 1	Organelle molecular trafficking	MRX41 (NS)
*300774	*RAB39B*	Xq28	Ras associated protein	Organelle molecular trafficking	MRX72 (NS)
*300019	*HCFC1*	Xq28	Host cell factor C1	Controlling of cell cycle	MRX3 (NS)
*300138	*CLIC2*	Xq28	Chloride intracellular channel 2	Regulation of calcium homeostasis in cardiac cells	MRXS32 (S)
*300005	*MECP2*	Xq28	Methyl-CpG-binding protein 2	Neuron maturation	Rett syndrome; MRXS13 (S); MRX (Lubs type) (S)

As shown in table 2, some of the XLID genes are involved in essential biological pathways and some are directly involved in synaptic activity and development of the nervous system. Therefore, it is plausible that alterations in the copy number of these genes will have moderate to severe consequences, depending on the number of genes affected.

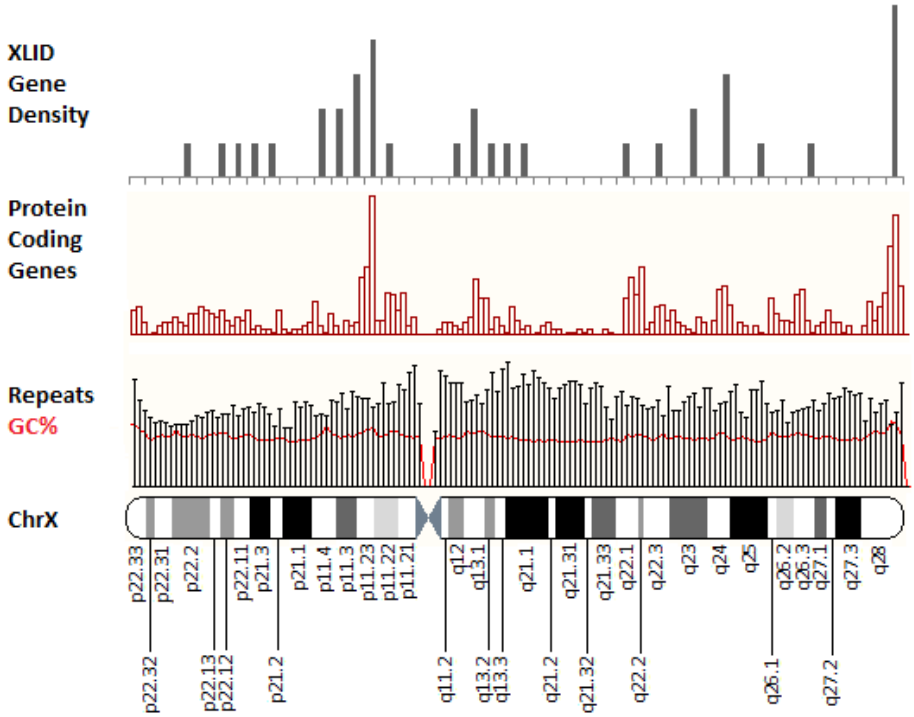


Fig. 6 – Density of repetitive elements, protein coding genes and the selected XLID-associated genes (n=37)[4].

Fig. 6 displays the XLID genes density on the X chromosome and reveals the hotspots. Despite the bias due to the sample size, it is worth mentioning that the "hotspots" correspond to previously reported unstable regions of the X chromosome. For instance, the Xq28 arm is an unstable region associated with several neurodevelopmental disorders and syndromes (e.g. Xq28 microduplication syndrome, Rett syndrome, etc.). The Xp11.23-Xp11.22 region is prone to recombination events and typically included within a Xp11.2 duplication associated with intellectual disability [56].

Search for pathogenic copy number variations associated with XLID

The search for pathogenic variants on DECIPHER v9.10 revealed 51 CNVs from distinct patients and different patterns of inheritance. Possibly, likely and definitely pathogenic variants are associated with distinct probabilities regarding the causality of XLID.

[4] Data retrieved from the **Ensembl** database (GRCh37/hg19 assembly) (http://grch37.ensembl.org/Homo_sapiens) [52]

Table 3 – Pathogenic copy number variants associated with X-linked intellectual disability genes.

Patient ID (DECIPHER v9.10)	**CNV**	**Inheritance / Pathogenicity**	**Type**	**Breakpoint coordinates (bp)** (GRCh37/hg19)	**Associated XLID genes**
288745, 289436, 289986	**1**	De novo constitutive/Possibly pathogenic	Gain	**X:60,691-155,246,653**	All 37 genes
288698	**2**	Unknown/Possibly pathogenic	Loss	**X:162,424-58,482,394**	*PTCHD1, HUWE1, USP9X, CLCN4, IQSEC2, FTSJ1, KDM5C, ZNF81, PQBP1, KLF8, ARX, TSPAN7, ZNF41, IL1RAPL1, RPS6KA3, ZNF674*
289320	**3**	De novo constitutive/Possibly pathogenic	Gain	**X:166,304-13,951,437**	*CLCN4*
288815	**4**	Unknown/Definitely pathogenic	Gain	**X:166,304-155,198,515**	All 37 genes
289471	**5**	Unknown/Possibly pathogenic	Gain	**X:166,304-155,208,397**	All 37 genes
275233	**6**	De novo constitutive/Definitely pathogenic	Del	**X:27,546,626-37,473,947**	*IL1RAPL1*
288399	**7**	Maternally inherited, constitutive in mother/Possibly pathogenic	Loss	**X:29,538,902-29,637,678**	*IL1RAPL1*
328709	**8**	Paternally inherited, constitutive in father/Possibly pathogenic	Del	**X:29,730,735-29,812,525**	*IL1RAPL1*
289703	**9**	Unknown/Possibly pathogenic	Gain	**X:29,871,994-30,082,735**	*IL1RAPL1*
314667	**10**	Unknown/Probably pathogenic	Del	**X:29,934,764-30,026,237**	*IL1RAPL1*
290238	**11**	Maternally inherited, constitutive in mother/Possibly pathogenic	Del	**X:38,398,928-38,582,588**	*TSPAN7*
289121, 289866, 290096, 290288	**12**	Paternally inherited, constitutive in father/Possibly pathogenic	Gain	**X:38,488,581-38,621,627**	*TSPAN7*
314823	**13**	Unknown/Possibly pathogenic	Dup	**X:38,488,637-38,547,932**	*TSPAN7*
332003	**14**	Unknown/Probably pathogenic	Del	**X:38,511,318-38,676,376**	*TSPAN7*
290181	**15**	Unknown/Possibly pathogenic	Loss	**X:47,330,199-47,334,872**	*ZNF41*
288574, 289575, 290210	**16**	Unknown/Possibly pathogenic	Loss	**X:47,330,199-47,335,098**	*ZNF41*
289561	**17**	Unknown/Possibly pathogenic	Gain	**X:47,414,181-48,204,109**	*ZNF81*
286242	**18**	De novo constitutive/Definitely pathogenic	Dup	**X:48,141,481-52,825,564**	*FTSJ1, PQBP1*
289849	**19**	Unknown/Possibly pathogenic	Gain	**X:48,204,030-52,624,130**	*FTSJ1, PQBP1*
289272	**20**	Maternally inherited, constitutive in mother/Possibly pathogenic	Loss	**X:48,289,368-48,349,670**	*FTSJ1*
289499	**21**	Unknown/Possibly pathogenic	Gain	**X:48,289,368-52,693,971**	*FTSJ1, PQBP1*
322832	**22**	De novo constitutive/Definitely pathogenic	Dup	**X:48,434,450-51,488,014**	*PQBP1*

291036	**23**	Maternally inherited, constitutive in mother/Definitely pathogenic	Dup	**X:53,228,159-54,133,745**	*PHF8, HUWE1, IQSEC2, KDM5C*
291035	**24**	Maternally inherited, constitutive in mother/Definitely pathogenic	Dup	**X:53,316,246-54,074,268**	*PHF8, HUWE1, IQSEC2*
289296	**25**	Maternally inherited, constitutive in mother/Possibly pathogenic	Gain	**X:53,453,926-53,713,965**	*HUWE1*
(*)	**26**	Pathogenic	Dup	**X:66,941,778-67,729,969**	*OPHN1*
288606	**27**	Unknown/Possibly pathogenic	Loss	**X:67,394,080-67,433,649**	*OPHN1*
(*)	**28**	Pathogenic	Del	**X:69,687,275-69,728,100**	*DLG3*
289249, 288224, 288694, 289428	**29**	Maternally inherited, constitutive in mother/Possibly pathogenic	Gain	**X:74,462,973-74,651,635**	*ZDHHC15*
283470	**30**	Unknown/Probably pathogenic	Dup	**X:74,494,004-74,649,820**	*ZDHHC15*
314839	**31**	Unknown/Possibly pathogenic	Dup	**X:79,915,497-80,007,002**	*BRWD3*
290295	**32**	Unknown/Possibly pathogenic	Loss	**X:79,952,082-79,978,501**	*BRWD3*
289519	**33**	Unknown/Definitely pathogenic	Loss	**X:93,755,145-100,251,990**	*PCDH19*
287183	**34**	De novo constitutive/Definitely pathogenic	Dup	**X:100,809,070-111,920,771**	*MID2, ACSL4, PAK3*
258495	**35**	De novo constitutive/Probably pathogenic	Del	**X:107,010,891-107,143,747**	*MID2*
327139	**36**	Unknown/Definitely pathogenic	Dup	**X:118,869,558-123,316,206**	*UPF3B, THOC2, CUL4B*
290060	**37**	De novo constitutive/Possibly pathogenic	Gain	**X:119,252,936-126,163,222**	*THOC2, CUL4B*
300665	**38**	Maternally inherited, constitutive in mother/Definitely pathogenic	Dup	**X:122,207,979-123,357,995**	*THOC2*
285178	**39**	Unknown/Probably pathogenic	Dup	**X:122,497,089-122,869,869**	*THOC2*
289207	**40**	Maternally inherited, constitutive in mother/Possibly pathogenic	Gain	**X:135,335,340-135,810,683**	*ARHGEF6*
323519	**41**	Imbalance arising from a balanced parental rearrangement/Definitely pathogenic	Gain	**X:148,094,889-155,208,254**	*MECP2, GDI1, CLIC2, RAB39B, HCFC1*
288130	**42**	Unknown/Possibly pathogenic	Gain	**X:153,291,208-153,379,847**	*MECP2*
290190	**43**	Maternally inherited/Possibly pathogenic	Gain	**X:153,505,460-153,871,986**	*GDI1*
(**)	**44**	Unknown	Gain	**X:153,564,843-153,882,630**	*GDI1*
(**)	**45**	Unknown	Gain	**X:153,564,843-153,889,019**	*GDI1*
314984	**46**	Maternally inherited, constitutive in mother/Possibly pathogenic	Dup	**X:153,576,880-153,832,734**	*GDI1*
307704	**47**	Maternally inherited, constitutive in mother/Possibly pathogenic	Dup	**X:153,576,905-153,832,705**	*GDI1*
287173	**48**	Unknown/Possibly pathogenic	Dup	**X:153,589,508-154,002,457**	*GDI1*

289986	**49**	Unknown/Possibly pathogenic	Gain	**X:153,664,474-153,666,574**	*GDI1*
288134	**50**	Unknown/Possibly pathogenic	Gain	**X:154,112,938-154,560,384**	*CLIC2, RAB39B*
288462	**51**	Paternally inherited, constitutive in father/Possibly pathogenic	Loss	**X:154,118,578-154,560,384**	*CLIC2, RAB39B*

(*) Variants retrieved from Isrie et al (2012) [37] and breakpoints remapped from the NCB36/hg18 assembly to the GRCh37/hg19 assembly

(**) Variants retrieved from Vandewalle et al (2009) [53]

As shown in Table 3, the majority of the selected 51 CNVs include at least one of the selected XLID genes.

Some of the selected XLID-associated variants in the X chromosome co-express with likely benign, potentially or definitely pathogenic variants located in other chromosomes (Appendix I). For the analysis of co-variants, only potentially or causative pathogenic variants were considered due to a possible influence on the ID phenotypic manifestations. For instance, one of the individuals that manifested the XLID phenotype and expressed the CNV-1 variant also revealed a 2.2 kb deletion on the Cri du Chat locus (chromosome 5) that could be associated with severe developmental delay. As such, the ID phenotype could also be related to this variant. However, it is important to remark that other patients expressing the same variant (CNV-1) and a XLID phenotype did not reveal a pathogenic co-variant.

The patient that carried variant 41 (CNV-41) also revealed a causative pathogenic variant on chromosome 9 that involved a considerable amount of developmental disorders related genes. Therefore, it is unclear if the ID phenotype is a consequence of the pathogenic X chromosome variant, the co-variant or both variants. The same inconclusive statement can be attributed to variant 17 (CNV-17) and its co-variant on chromosome 11.

A patient with the possibly pathogenic CNV-49 variant manifested three distinct variants on chromosome 6 and chromosome 22. The loci of two of these variants are associated with syndromes that cause developmental delay (22q11 duplication syndrome and Phelan-McDermid syndrome); therefore it is unclear if the ID phenotype is a consequence of the X-CNV. The CNV-49 variant (2 kb) is located within the Xq28 duplication syndrome locus (257 kb) and only involves one dosage-sensitive gene (*GDI1*) that is highly expressed in the brain and is associated with XLID [53].

The observations indicate that analyses of the variants that co-occur with the X-linked CNV are still important when the geneticist attempts to establish genotype-phenotype relationships.

Repeat masking and proportions of repetitive elements

The analysis of repetitive elements within the breakpoints of the selected pathogenic CNVs revealed the prevalence of LINE elements (≈31.0%) and SINE elements (≈21.4%). The long terminal repeats (LTRs), DNA elements and other elements represent 14.4%, 13.5% and 19.7% of the total repeats on the 51 CNVs, respectively. The coverage of each element within its corresponding category is shown in Fig. 7.

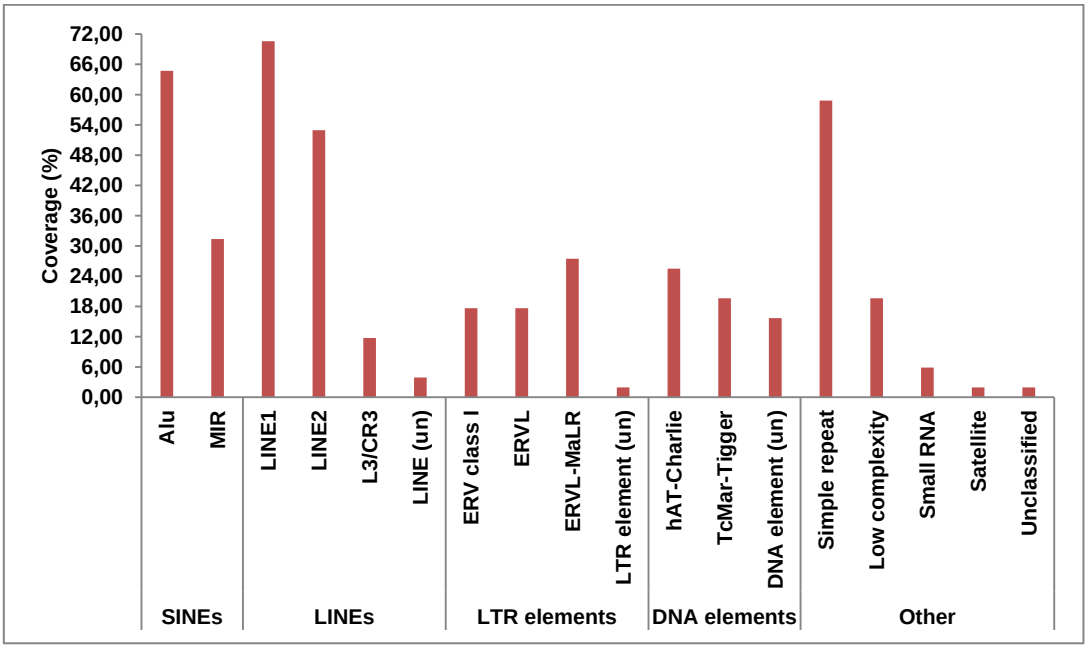


Fig. 7 – Frequency of five categories of repetitive elements present in the 51 XLID-associated CNVs.

In general LINE1 are the most abundant repeats (70.6%), followed by Alu (64.7%), simple repeats (58.8%), LINE2 (52.9%), ERVL-MaLR (27.5%), MIRs (31.4%), hAT-Charlie (25.5%), TcMar-Tigger (19.6%), low complexity repeats (19.6%), ERV class I (17.7%), ERVL (17.7%), L3/CR1 (11.8%), small RNAs (5.9%), satellites (2.0%) and at last, some unclassified elements which in total represent 2.0% (Fig. 7).

Within the SINEs category, about 67.3% of these elements represent Alu and 32.7% MIR. Alu elements are considered one of the most successful class of retrotransposons in the human genome and they account for 11% of the human genome [34]. The result of this breakpoint analysis revealed that Alu elements are in fact very abundant in the genome and that they are the second most abundant class of retrotransposons flanking the breakpoints of pathogenic or potentially pathogenic copy number variants. Overall, these elements are naturally abundant in the human genome; however it is important to underline that they may generate microdeletions through Alu-Alu recombination events. Therefore, it is quite plausible that some of the presented smaller variants with a high Alu density nearby the breakpoints suffered recombination between two or more homologous Alus.

The LINEs category included the general and categorical most abundant repeats L1 (50.7%), LINE2 (38.0%), the less abundant L3/CR1 (8.5%) and some unclassified elements (2.8%), which probably represent L1 or LINE2.

The autonomous L1 elements are the most abundant transposons in mammals and together with the Alu elements constitute the exclusive active transposons in humans. Despite the fact that L1 elements are very abundant, only a small amount of copies (around 80-100) retain their retrotransposition ability since most of their copies are actually defective and inactive [29, 33]. The LINE2 elements also represent a great portion of the LINEs category. However, both LINE2 elements and L3/CR1 elements represent inactive TEs [33].

Within the LTR elements, ERV-like sequences with mammalian long terminal repeats (ERVL-MaLR) are the dominant repeats (42.4%), followed by 27.3% of ERV class I elements and 27.3% of ERVLs. About 3.0% of the LTR elements remained unclassified.

The DNA elements category included DNA transposons such as hAT-Charlie, which represents 41.9% of these elements, TcMar-Tigger (32.3%) and some unclassified elements (25.6%).

Other elements detected by **RepeatMasker** and classified as a category were the abundant simple repeats (66.7%), low complexity repeats (22.2%), small RNA (6.7%), satellites (2.2%) and some other unclassified repeats (2.2%).

Overall, these results suggest that LINE1, Alu, simple repeats and LINE2 constitute the majority of the repetitive sequences within the flanking regions of the breakpoints. Both SINEs and LINEs are associated with insertion mutations, human disease and potential involvement in pseudogene formation [57]. Furthermore, Alu repeats are preferentially incorporated in genomic regions with high GC content, which points to the proximity to functional genes [20], suggesting influence on nearby genes. However, Alus are also very frequent in terms of the human genome, occurring with a rate of more than once every 3 kb, so it is expected that such a large portion of these elements are present within 4 kb, which is the total sequence size of the analyzed flanking regions of the CNV breakpoints. As stated before, LINEs and in particular L1 are very abundant in the X chromosome [40, 41] and these results clearly show this fact since L1 constitute approximately 71% of the analyzed sequences.

The fact that both LINE1 and Alu preserve retrotransposition activity and represent the most abundant elements within a 1 kb radius from the breakpoints of pathogenic CNVs suggests their influence on the structural instability.

Analysis of repeats overlapping the pathogenic breakpoints

In order to further investigate and characterize the precise genomic architecture of breakpoints of the pathogenic CNVs, the detailed region of each breakpoint was analyzed on UCSC Genome Browser. Repetitive elements of distinct features and families were found in the exact location of some of the breakpoints (Appendix II).

The analysis revealed that some breakpoints of XLID variants lie within one or more repetitive sequences. About 12.7% of the breakpoints (n=102) of the 51 XLID-associated CNVs were lying on a retrotransposon sequence. As would be expected, LINEs were the most frequent repeats. Intra-chromosomal segmental duplications that frequently generate local instability were found to incorporate some of the pathogenic breakpoints.

It is however important to note that the majority of the selected variants are non-recurrent and thus estimating their underlying mechanisms of formation is challenging and would require extensive research. Homology may have played a role in the formation of these variants however as previously stated homology is not implicit in DNA repair mechanisms such as non-homologous end joining (NHEJ), which does not

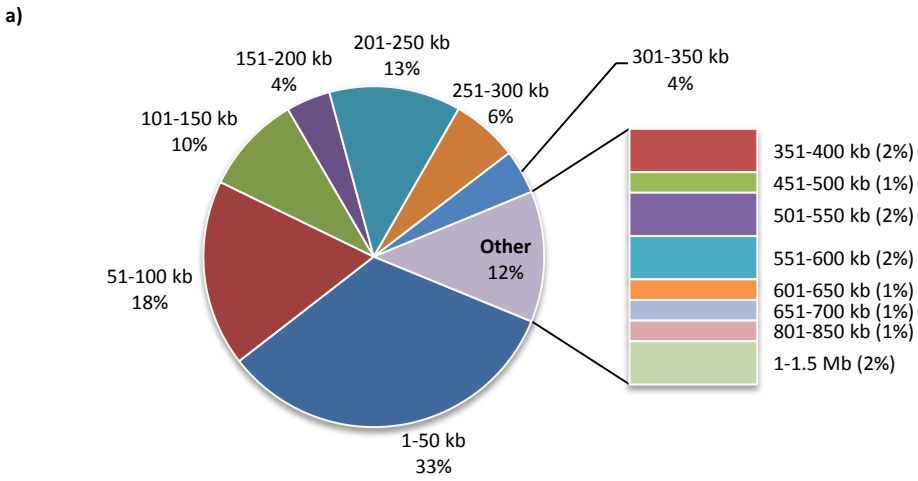
require substrate homology and may involve the joining of structural variants of several base pairs [19].

Thus, although it is very unlikely that these variants resulted from NAHR events it is plausible that similar putative variants with different sizes may rise from homology-based recombination events due to an abundance of intra-chromosomal homologous repeats.

About 15.0% of the breakpoints (n=40) of the 20 pathogenic co-variants were found within a repetitive sequence. Despite the small sample size that may bias the results, this combined information shows that pathogenic breakpoints may lie within repetitive sequences that lead to genomic instability and therefore may trigger the formation of new variants.

Population genetics analysis

In order to visualize the minimum distance between the breakpoints of a XLID-associated variant and a non-disease associated CNV reported by Sudmant et al. (2015) [8] we plotted the corresponding distances in Fig. 8.



b)

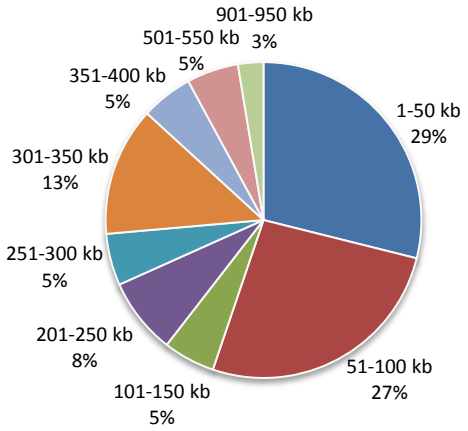


Fig. 8– Sequence size separating the breakpoints of pathogenic XLID-associated CNVs (N=51) and the breakpoints of non-pathogenic CNVs (N=43) **(a)** and sequence size separating the breakpoints of pathogenic co-variants of XLID-associated CNVs (N=20) and the breakpoints of non-pathogenic CNVs (N=30) **(b)**.

The 43 non-disease associated CNVs proximal to the breakpoints of the 51 pathogenic CNVs revealed a minimum distance of 0.8 kb and the maximum distance of 1.3 Mb. Overall, 33% of the sequence sizes between a pathogenic breakpoint and a non-pathogenic breakpoint were around 1-50 kb whereas sequence sizes longer than 50 kb and shorter than 100 kb represented 18% (Fig. 8a). Categories above 251 kb individually represent approximately or less than 10%. Within the 1-50 kb category, 34% of the breakpoints were less than 10 kb apart, suggesting that the majority of non-pathogenic CNVs are very proximal to pathogenic CNVs. The second most prevalent category was 31-40 kb (22%), followed by 41-50 kb (19%), 21-30 kb (16%) and finally 11-20 kb (9%).

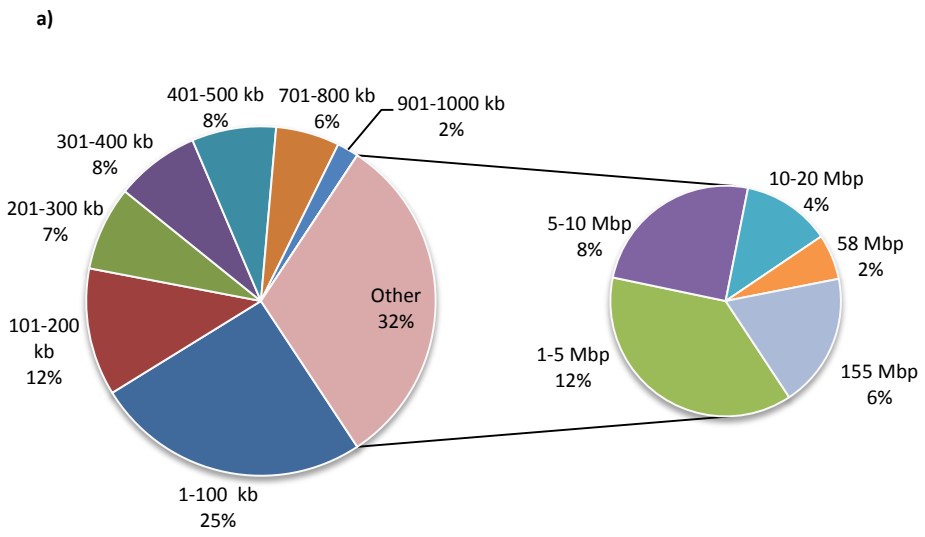
These results suggest proximity between non-pathogenic variants and pathogenic variants, therefore underlining the complexity of copy number variants. Variants that differ in a few nucleotides can affect individuals in distinct ways; one CNV may be involved in normal variation among population whereas the other may cause clinical phenotypic manifestations. For instance, a non-pathogenic variant was reported only 800 nucleotides apart from pathogenic variant CNV-11 which is associated with an autistic and intellectual disability phenotype. However, CNV-11 is a 184 kb deletion overlapping a XLID gene (*TSPAN7*) and a pseudogene whereas the non-pathogenic CNV is a 7 kb deletion that does not overlap any gene.

Regarding the pathogenic co-variants of the selected XLID-associated CNVs, 30 non-pathogenic CNVs were selected for the analysis due to their proximity to the pathogenic breakpoints. Similarly to X-CNVs, most breakpoints of the pathogenic co-variants are located in the vicinities of non-pathogenic CNVs (1-100 kb apart) (Fig. 8b). About 29% of the sequence size separating both types is 1-50 kb, with a considerable proportion (27%) falling within the ≤ 10 kb category. Once again, the complexity of copy number variants is underlined in this analysis. In this co-variant analysis, the shortest noted distance between both types of breakpoints was 2 kb whereas the longest was 914 kb.

This analysis revealed than many non-pathogenic CNVs intersect both types of pathogenic CNVs and are within their sequence, consequently challenging the understanding of the pathogenicity of a structural variant. It is likely that even though these non-pathogenic CNVs intersect pathogenic variants, their sequences are comparatively small and do not include promoter regions or biologically important genes.

Copy number variant size

A comparative analysis between the sizes of both types of CNVs was performed to validate the influence of the variant size in pathogenicity.



b)

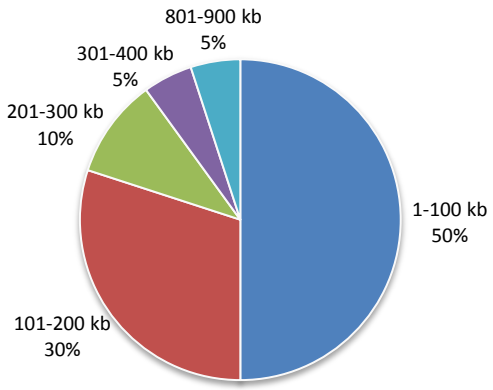


Fig. 9 – Size of the selected XLID-associated pathogenic CNVs (n=51). Sizes of ≥ 1000 kb (1Mbp) are represented in the isolated section **(a).** Size of pathogenic CNVs (n=20) co-segregated with the selected XLID-associated pathogenic CNVs **(b).**

Regarding CNV size, the majority of the selected XLID-associated CNVs were found to be 1-100 kb in size (25%) (Fig. 9a). Furthermore, 32% of these variants are very large in size (≥ 1 Mbp) and 6% of these CNVs encompass the entire X chromosome sequence (155 Mbp) which includes many neurodevelopmental genes. The fact that in such a small sample size (n=51) around 32% of the CNVs are very large suggests that variant size may be an important factor influencing pathogenicity. It would be of interest to perform a more exhaustive analysis with a larger sample to confirm this influence.

Additionally, the same pattern was noted regarding the size of pathogenic co-variants segregating with the XLID-associated CNVs (Fig. 9b). The majority of variants were found to be 1-100 kb in size (50%) and similarly to the XLID variants, the second most frequent category was 101-200 kb (30%). Pathogenic variants greater than 1 Mbp in size were not reported for the co-variants.

In conclusion, pathogenic CNVs associated with XLID genes tend to either have a small to moderate size (1-200 kb, 37%) or a very large size (1-155 Mbp, 32%).

A comparative analysis between the sizes of the different copy number variants revealed that pathogenic CNVs tend to be larger in size (Table 4).

Table 4 - Comparative analysis between sizes of non-pathogenic CNVs (X chromosome and autosomes) and pathogenic CNVs (selected XLID-associated variants and their co-variants).

	Coverage			
	Pathogenic CNVs		**Non-pathogenic CNVs**	
Size	**XLID variants (n=51)**	**Co-variants (n=20)**	**X chromosome (n=43)**	**Autosomes (n=30)**
1-50 kb	11.8%	40.0%	88.4%	83.3%
51-100 kb	13.7%	10.0%	2.3%	10.0%
>100 kb	74.5% (42.0% ≥ 1Mbp)	50.0%	9.3%	6.7%

While non-pathogenic CNVs are mostly 1-50 kb long, most pathogenic variants are more than 100 kb in size. Particularly, 42% of the XLID variants with more than 100 kb are larger than 1 Mbp which considerably affects the genomic sequence. These discrepancies between CNV sizes are interesting since larger variants tend to affect a wider group of genes and consequently cause disease.

As shown in table 2, chromosome X encodes crucial genes involved in neuronal development and maturation, synaptic signaling pathways, hedgehog signaling pathways (required for embryonic development), transcription regulation, cell stabilization, DNA repair mechanisms and biochemical pathways involving fatty acids that constitute myelin sheaths. Therefore, it is important to remark that copy number variants influencing these genes, whether by dosage malfunction or interference in gene regulation, will likely cause severe damage and influence embryonic development. However, it is still important to remark that smaller variants might still be equally damaging depending on the affected genomic regions.

Conclusions

Copy number variants are useful but challenging molecular markers due to their structural complexity and different levels of phenotypic penetrance among individuals, which results in the difficulty in estimating the frequency of pathogenic variants.

Repetitive elements such as low copy repeats and high copy repeats are deeply involved in the formation of these variants and may lead to complex rearrangements that may manifest in the form of a clinical phenotype. Requirements for the occurrence of NAHR events between highly homologous sequences vary among individuals; consequently, the frequency estimation for a given pathogenic CNV that resulted from NAHR is challenging. Additionally, less frequent mechanisms (e.g., NHEJ, FoSTeS, BIR, etc.) may lead to CNV formation, which in turn adds to the complexity of these variants. Whenever a new pathogenic variant is reported several mechanistic ways of formation are proposed and hypothesized. However, only a few are confirmed. Crucial facts such as hereditability (maternally/paternally inherited, *de novo* variant) and pathogenicity level are missing in most of the reported variants on the online databases.

Genomic sequencing and *in silico* analysis of the flanking regions of pathogenic breakpoints can lead to the detection and better understanding of the unstable genomic regions (recombination hotspots) that predispose to these events. If homologous repetitive elements are scarce in a specific region, then more complex mechanisms (e.g., errors of DNA damage repair mechanisms, etc) may have been involved.

Despite size being an influencing factor on pathogenicity, small variants can also cause severe consequences and lead to profound genomic disorders and syndromes. Moreover, these variants are less likely to involve complex molecular mechanisms but instead result of a simple recombination event between proximal homologous repeats such as LCRs, LINEs or Alus, which are the most abundant repeats.

Regarding the case of X-linked intellectual disability, copy number variants tend to behave similarly to other markers. The involvement of repetitive elements and the specific mechanisms of XLID-CNV formation are unknown and hard to predict since some CNVs seem to be influenced by the repeats and others do not. Some CNVs may have been originated through NAHR events between homologous repeats whereas other CNVs, particularly non-recurrent variants [18], might have been involved in more

complex mechanisms that do not specifically require repeats (e.g., NHEJ, FoSTeS, BIR, etc.).

Additionally, most pathogenic XLID CNVs are very proximal to non-pathogenic CNVs on the X chromosome and there is a considerable discrepancy between their sequence sizes. XLID-associated pathogenic copy number variants tend to be larger in size than non-pathogenic CNVs, therefore affecting more genes. Some pathogenic variants involve most of the official XLID genes, which are scattered throughout the X chromosome and one particular variant of 155 Mbp was reported in three different patients.

It would be interesting to replicate this particular analysis with a larger sample size or a different phenotype to further investigate the correlation between pathogenic and non-pathogenic copy number variants whenever more data becomes available in the literature. It would also be of interest to do an extensive research and detailed characterization of the elements that predispose the formation of structural variants in individuals. Currently, some factors are known to influence NAHR events however there are still some additional unknown influencing factors that could be very important for clinical genetics. The same applies to other molecular mechanisms that lead to CNV formation. More important than the size of the CNV itself is the mechanism behind its formation and hereditability pattern because it could help to unravel its pathogenicity.

Although copy number variants are still not widely understood and many questions raise regarding these structural mutations, it still is a very promising and relevant topic of research. In the future, a complete database of all the compiled data from exhaustive population studies that have been regularly performed throughout the years will be of great interest and it will help many researchers and clinicians.

Bibliographic References

1. Henrichsen, C.N., E. Chaignat, and A. Reymond, *Copy number variants, diseases and gene expression.* Hum Mol Genet, 2009. **18**(R1): p. R1-8.
2. Freeman, J.L., et al., *Copy number variation: new insights in genome diversity.* Genome Res, 2006. **16**(8): p. 949-61.
3. Itsara, A., et al., *Population analysis of large copy number variants and hotspots of human genetic disease.* Am J Hum Genet, 2009. **84**(2): p. 148-61.
4. Radke, D.W. and C. Lee, *Adaptive potential of genomic structural variation in human and mammalian evolution.* Brief Funct Genomics, 2015. **14**(5): p. 358-68.
5. Jobling, M.A., et al., *Human Evolutionary Genetics*. 2nd ed. 2014, New York: Garland Science.
6. Veerappa, A.M., et al., *Global spectrum of copy number variations reveals genome organizational plasticity and proposes new migration routes.* PLoS One, 2015. **10**(4): p. e0121846.
7. Zarrei, M., et al., *A copy number variation map of the human genome.* Nat Rev Genet, 2015. **16**(3): p. 172-83.
8. Sudmant, P.H., et al., *Global diversity, population stratification, and selection of human copy-number variation.* Science, 2015. **349**(6253): p. aab3761.
9. Ewald, I.P., et al., *Genomic rearrangements in BRCA1 and BRCA2: A literature review.* Genet Mol Biol, 2009. **32**(3): p. 437-46.
10. Lee, J.A. and J.R. Lupski, *Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders.* Neuron, 2006. **52**(1): p. 103-21.
11. Bosch, E. and M.A. Jobling, *Duplications of the AZFa region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility.* Hum Mol Genet, 2003. **12**(3): p. 341-7.
12. Saunier, S., et al., *Characterization of the NPHP1 locus: mutational mechanism involved in deletions in familial juvenile nephronophthisis.* Am J Hum Genet, 2000. **66**(3): p. 778-89.
13. Ionita-Laza, I., et al., *Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis.* Genomics, 2009. **93**(1): p. 22-6.
14. Bassaganyas, L., et al., *Worldwide population distribution of the common LCE3C-LCE3B deletion associated with psoriasis and other autoimmune disorders.* BMC Genomics, 2013. **14**: p. 261.

15. Conrad, D.F., et al., *Origins and functional impact of copy number variation in the human genome.* Nature, 2010. **464**(7289): p. 704-12.
16. Picanco, J.B., et al., *Tri-allelic pattern at the TPOX locus: a familial study.* Gene, 2014. **535**(2): p. 353-8.
17. Repnikova, E.A., et al., *Characterization of copy number variation in genomic regions containing STR loci using array comparative genomic hybridization.* Forensic Sci Int Genet, 2013. **7**(5): p. 475-81.
18. Carvalho, C.M. and J.R. Lupski, *Mechanisms underlying structural variant formation in genomic disorders.* Nat Rev Genet, 2016. **17**(4): p. 224-38.
19. Chen, L., et al., *Genome architecture and its roles in human copy number variation.* Genomics Inform, 2014. **12**(4): p. 136-44.
20. Strachan, T. and A. Read, *Human Molecular Genetics*. 4th ed. 2011, New York: Garland Science
21. Peng, Z., et al., *Correlation between frequency of non-allelic homologous recombination and homology properties: evidence from homology-mediated CNV mutations in the human genome.* Hum Mol Genet, 2015. **24**(5): p. 1225-33.
22. Dittwald, P., et al., *NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits.* Genome Res, 2013. **23**(9): p. 1395-409.
23. Gu, W., F. Zhang, and J.R. Lupski, *Mechanisms for human genomic rearrangements.* Pathogenetics, 2008. **1**(1): p. 4.
24. Campbell, I.M., et al., *Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination.* BMC Biol, 2014. **12**: p. 74.
25. Stankiewicz, P. and J.R. Lupski, *Genome architecture, rearrangements and genomic disorders.* Trends Genet, 2002. **18**(2): p. 74-82.
26. Weckselblatt, B. and M.K. Rudd, *Human Structural Variation: Mechanisms of Chromosome Rearrangements.* Trends Genet, 2015. **31**(10): p. 587-99.
27. Wicker, T., et al., *A unified classification system for eukaryotic transposable elements.* Nat Rev Genet, 2007. **8**(12): p. 973-82.
28. Smit, A.F. and A.D. Riggs, *Tiggers and DNA transposon fossils in the human genome.* Proc Natl Acad Sci U S A, 1996. **93**(4): p. 1443-8.
29. Richardson, S.R., et al., *The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes.* Microbiol Spectr, 2015. **3**(2): p. MDNA3-0061-2014.
30. Smit, A.F., *Identification of a new, abundant superfamily of mammalian LTR-transposons.* Nucleic Acids Res, 1993. **21**(8): p. 1863-72.

31. Smit, A.F., *The origin of interspersed repeats in the human genome.* Curr Opin Genet Dev, 1996. **6**(6): p. 743-8.
32. Munoz-Lopez, M. and J.L. Garcia-Perez, *DNA transposons: nature and applications in genomics.* Curr Genomics, 2010. **11**(2): p. 115-28.
33. Ostertag, E.M. and H.H. Kazazian, Jr., *Biology of mammalian L1 retrotransposons.* Annu Rev Genet, 2001. **35**: p. 501-38.
34. Deininger, P., *Alu elements: know the SINEs.* Genome Biol, 2011. **12**(12): p. 236.
35. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
36. Arensburger, P., et al., *Phylogenetic and functional characterization of the hAT transposon superfamily.* Genetics, 2011. **188**(1): p. 45-57.
37. Isrie, M., et al., *Sporadic male patients with intellectual disability: contribution of X chromosome copy number variants.* Eur J Med Genet, 2012. **55**(11): p. 577-85.
38. Delbridge, M.L., et al., *Origin and evolution of candidate mental retardation genes on the human X chromosome (MRX).* BMC Genomics, 2008. **9**: p. 65.
39. Singh, N.D. and D.A. Petrov, *Evolution of gene function on the X chromosome versus the autosomes.* Genome Dyn, 2007. **3**: p. 101-18.
40. Ross, M.T., et al., *The DNA sequence of the human X chromosome.* Nature, 2005. **434**(7031): p. 325-37.
41. Lyon, M.F., *LINE-1 elements and X chromosome inactivation: a function for "junk" DNA?* Proc Natl Acad Sci U S A, 2000. **97**(12): p. 6248-9.
42. Tarpey, P.S., et al., *A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation.* Nat Genet, 2009. **41**(5): p. 535-43.
43. Hehir-Kwa, J.Y., et al., *Pathogenic or not? Assessing the clinical relevance of copy number variants.* Clin Genet, 2013. **84**(5): p. 415-21.
44. Leonard, H. and X. Wen, *The epidemiology of mental retardation: challenges and opportunities in the new millennium.* Ment Retard Dev Disabil Res Rev, 2002. **8**(3): p. 117-34.
45. Gecz, J., C. Shoubridge, and M. Corbett, *The genetic landscape of intellectual disability arising from chromosome X.* Trends Genet, 2009. **25**(7): p. 308-16.
46. Piton, A., C. Redin, and J.L. Mandel, *XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing.* Am J Hum Genet, 2013. **93**(2): p. 368-83.

47. Greenwood Genetic Center. *XLID*. [cited 2015 September 1]; Available from: http://www.ggc.org/research/molecular-studies/xlid.html.
48. van Bokhoven, H., *Genetic and epigenetic networks in intellectual disabilities.* Annu Rev Genet, 2011. **45**: p. 81-104.
49. Van Esch, H., et al., *Duplication of the MECP2 region is a frequent cause of severe mental retardation and progressive neurological symptoms in males.* Am J Hum Genet, 2005. **77**(3): p. 442-53.
50. Moyses-Oliveira, M., et al., *X-linked intellectual disability related genes disrupted by balanced X-autosome translocations.* Am J Med Genet B Neuropsychiatr Genet, 2015. **168**(8): p. 669-77.
51. Firth, H.V., et al., *DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources.* Am J Hum Genet, 2009. **84**(4): p. 524-33.
52. Yates, A., et al., *Ensembl 2016.* Nucleic Acids Res, 2016. **44**(D1): p. D710-6.
53. Vandewalle, J., et al., *Dosage-dependent severity of the phenotype in patients with mental retardation due to a recurrent copy-number gain at Xq28 mediated by an unusual recombination.* Am J Hum Genet, 2009. **85**(6): p. 809-22.
54. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
55. McKusick-Nathans Institute of Genetic Medicine. *Online Mendelian Inheritance in Man, OMIM*. [cited 2016 August 11]; Available from: http://omim.org/.
56. Froyen, G., et al., *Copy-number gains of HUWE1 due to replication- and recombination-based rearrangements.* Am J Hum Genet, 2012. **91**(2): p. 252-64.
57. Giordano, J., et al., *Evolutionary history of mammalian transposons determined by genome-wide defragmentation.* PLoS Comput Biol, 2007. **3**(7): p. e137.

Appendices

Appendix I: Co-variants expressed with the selected X-CNVs.

Appendix II: Repetitive elements overlapping some of the breakpoints of XLID-associated CNVs and co-variants of XLID-associated CNVs.

Appendix I: Co-variants expressed with the selected X-CNVs[5].

Patient ID (DECIPHER v9.10)	X-CNV	Type	Co-variant coordinates (bp) (GRCh37/hg19)	Pathogenicity	ID related phenotype
289986	**1**	Del	**chr5**:16,487-235,903	Possibly pathogenic	*SDHA* is a DD gene
288698	**2**	Del	**chr2**:79,969,663-80,073,712	Possibly pathogenic	-
		Dup	**chr18**:19,075,518-19,344,079	Possibly pathogenic	-
289320	**3**	Del	**chrY**:16,841,656-59,349,659	Possibly pathogenic	-
288815	**4**	Del	**chr12**:24,136,874-24,295,420	Possibly pathogenic	-
290228	**12**	Del	**chr20**:9,211,722-9,343,865	Possibly pathogenic	*PLCB4* is a DD gene
290181	**15**	Dup	**chr8**:132,480,029-132,565,953	Possibly pathogenic	-
288574	**16**	Dup	**chr9**:260,360-416,483	Possibly pathogenic	*DOCK8* is a DD gene
290210		Dup	**chr19**:42,676,550-43,757,775	Possibly pathogenic	*ERF* and *MEGF8* are DD genes
289561	**17**	Del	**chr11**:77,251,642-79,935,761	Definitely pathogenic	*ALG8* is a DD gene
289499	**21**	Dup	**chr2**:2,581,387-3,584,779	Possibly pathogenic	-
288606	**27**	Dup	**chr3**:10,092,320-10,288,971	Possibly pathogenic	*FANCD2* is a DD gene
283470	**30**	Dup	**chr14**:26,658,905-27,538,941	Probably pathogenic	-
		Dup	**chr5**:58,885,448-59,212,327	Probably pathogenic	*PDE4D* is a DD gene
290295	**32**	Dup	**chr19**:366,719-382,893	Possibly pathogenic	-
323519	**41**	Del	**chr9:** 204,183-5,951,402	Definitely pathogenic	*DOCK8, GLIS3, MARCA2, VLDLR* are DD genes
289983	**49**	Del	**chr6**:56,627,396-56,695,989	Possibly pathogenic	-
		Dup	**chr22**:19,747,556-19,749,395	Possibly pathogenic	*TBX1* is a DD gene
		Del	**chr22**:51,161,423-51,172,097	Possibly pathogenic	*SHANK3* is a probable DD gene
288134, 288462	**50 & 51**	Del	**chr4**:184,999,811-185,119,826	Possibly pathogenic	-

[5] DD is an abbreviation for Developmental Disorder and indicates genes belonging to the Developmental Disorders Genotype-Phenotype Database (DDG2P) and confirmed as being associated with developmental disorders.

Appendix II: Repetitive elements overlapping some of the breakpoints of XLID-associated CNVs and co-variants of XLID-associated CNVs.

		Repeats lying on pathogenic breakpoints	
X-CNV	**Breakpoint coordinate (bp)** (GRCh37/hg19)	**Repeat type**	**Repeat coordinates (bp)** (GRCh37/hg19)
2	**chrX:162,424**	Simple tandem repeat	chrX:154,937-166,217
	chrX:58,482,394	L1PA3 (LINE)	chrX:58,482,386-58,482,785
6	**chrX:27,546,626**	TA(n) (Simple repeat)	chrX:27,546,540-27,546,659
8	**chrX:29,812,525**	LTR82B (LTR)	chrX:29,812,481-29,813,195
9	**chrX:30,082,735**	MamGypLTR2c (LTR)	chrX:30,082,643-30,082,774
15 & 16	**chrX:47,330,199**	L1MEA3 (LINE)	chrX:47,330,162-47,330,285
16	**chrX:47,334,872**	L1M5 (LINE)	chrX:47,334,868-47,335,006
23	**chrX:54,133,745**	AluSx (SINE)	chrX:54,133,521-54,133,824
24	**chrX:54,074,268**	L2a (LINE)	chrX:54,074,019-54,074,340
25	**chrX:53,453,926**	AluSx (SINE)	chrX:53,453,654-53,453,935
26	**chrX:67,729,969**	L1MCa (LINE)	chrX:67,729,693-67,730,074
36	**chrX:123,316,206**	AluSx (SINE)	chrX:123,316,140-123,316,223
39	**chrX:122,869,869**	Tigger10 (TcMar-Tigger)	chrX:122,869,636-122,869,910
43	**chrX:153,505,460**	L1MB5 (LINE)	chrX:153,505,356-153,505,484
		Segmental duplication	chrX:153,481,277-153,519,082
44 & 45	**chrX:153,564,843**	Segmental duplication	chrX:153,564,285-153,575,614
48	**chrX:154,002,457**	L2b (LINE)	chrX:154,002,320-154,002,633
50	**chrX:154,112,938**	Segmental duplication	chrX:154,109,089-154,118,602
51	**chrX:154,118,578**	L2b (LINE)	chrX:154,118,568-154,118,613
		Segmental duplication	chrX:154,109,089-154,118,602

		Repeats lying on pathogenic breakpoints	
X-CNV	**Co-variant breakpoint coordinate (bp)** (GRCh37/hg19)	**Repeat type**	**Repeat coordinates (bp)** (GRCh37/hg19)
1	**chr5**:16,487	L1MC3 (LINE)	chr5:16,473-17,482
		Segmental duplication	chr5:15,294-19,553
	chr5: 235,903	Segmental duplication	chr5:235,806-259,327
15	**chr8**:132,565,953	A-rich (Low complexity)	chr8:132,565,946-132,566,002
16	**chr19**:43,757,775	LTR16B2 (LTR)	chr19:43,757,753-43,758,149
		Segmental duplication	chr19:43,757,731-43,758,380
27	**chr3**:10,092,320	AluY	chr3:10,092,029-10,092,327
	chr3:10,288,971	AluJb	chr3:10,288,965-10,289,266